Proceedings of Machine Learning Research vol 134:1–44, 2021 34th Annual Conference on Learning Theory

Sparse sketches with small inversion bias

Michał Dereziński MDEREZIN@BERKELEY.EDU
Department of Statistics, UC Berkeley

Zhenyu Liao ZHENYU.LIAO@BERKELEY.EDU
ICSI and Department of Statistics, UC Berkeley

Edgar Dobriban DOBRIBAN@WHARTON.UPENN.EDU
Department of Statistics and Data Science, University of Pennsylvania

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics, UC Berkeley

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

For a tall $n \times d$ matrix $\mathbf{A}$ and a random $m \times n$ sketching matrix $\mathbf{S}$, the sketched estimate of the inverse covariance matrix $(\mathbf{A}^\top \mathbf{A})^{-1}$ is typically biased: $\mathbb{E}[(\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}})^{-1}] \neq (\mathbf{A}^\top \mathbf{A})^{-1}$, where $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$. This phenomenon, which we call *inversion bias*, arises, e.g., in statistics and distributed optimization, when averaging multiple independently constructed estimates of quantities that depend on the inverse covariance matrix. We develop a framework for analyzing inversion bias, based on our proposed concept of an (ϵ, δ)-unbiased estimator for random matrices. We show that when the sketching matrix $\mathbf{S}$ is dense and has i.i.d. sub-gaussian entries, then after simple rescaling, the estimator $(\frac{m}{m-d}\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}})^{-1}$ is (ϵ, δ)-unbiased for $(\mathbf{A}^\top \mathbf{A})^{-1}$ with a sketch of size $m = O(d + \sqrt{d}/\epsilon)$. In particular, this implies that for $m = O(d)$, the inversion bias of this estimator is $O(1/\sqrt{d})$, which is much smaller than the $\Theta(1)$ approximation error obtained as a consequence of the subspace embedding guarantee for sub-gaussian sketches. We then propose a new sketching technique, called LEverage Score Sparsified (LESS) embeddings, which uses ideas from both data-oblivious sparse embeddings as well as data-aware leverage-based row sampling methods, to get ϵ inversion bias for sketch size $m = O(d \log d + \sqrt{d}/\epsilon)$ in time $O(\mathrm{nnz}(\mathbf{A}) \log n + md^2)$, where nnz is the number of non-zeros. The key techniques enabling our analysis include an extension of a classical inequality of Bai and Silverstein for random quadratic forms, which we call the *Restricted Bai-Silverstein inequality*; and anti-concentration of the Binomial distribution via the Paley-Zygmund inequality, which we use to prove a lower bound showing that leverage score sampling sketches generally do not achieve small inversion bias.

Keywords: randomized linear algebra, sketching, random matrix theory, distributed optimization

1. Introduction

Sketching has been widely used in the design of scalable algorithms, perhaps most prominently in Randomized Numerical Linear Algebra (RandNLA) due to applications in machine learning and data analysis. In this approach, one randomly samples or computes a random projection of the data matrix to construct a smaller matrix, the sketch. One then uses the sketch as a surrogate to approximate quantities of interest. The analysis of these methods typically proceeds via a Johnson-Lindenstrauss-type argument to establish that the geometry of the matrix is not perturbed too much under the sketching operation. These methods have yielded state-of-the-art in worst-case analysis, high-quality numerical implementations, and numerous applications in machine learning (Mahoney, 2011; Halko et al., 2011b; Woodruff, 2014; Drineas and Mahoney, 2016, 2018).

In many cases, either to preserve the structure of the data or for algorithmic reasons, one is interested in sparse sketches, i.e., random transformations that are represented by matrices with most entries exactly equal to zero. One class of sparse sketches includes row sampling techniques, such as leverage score sampling (Drineas et al., 2008, 2012; Ma et al., 2015), which are typically *data-aware*, in the sense that the sampling distribution depends on the given data matrix. Another important class of methods uses *data-oblivious* sparse embedding, such as the CountSketch (Charikar et al., 2002; Clarkson and Woodruff, 2017; Meng and Mahoney, 2013; Nelson and Nguyên, 2013), to construct sketches in time depending on the number of non-zeros (nnz) in the input.

In all these cases, one can show that the sketch will be an approximation of the solution with high probability. However, comparatively little is known about the average performance of these sketches. In particular, there may be a systematic bias away from the solution, which is problematic in many situations in statistics, machine learning, and data analysis. Perhaps the most ubiquitous example of this phenomenon is the systematic bias caused by matrix inversion, a key component of algorithms in the aforementioned domains. In this paper, we introduce the fundamental notion of inversion bias, which provides a finer control over the sketched estimates involving matrix inversion. We show that one can conveniently make the inversion bias small with dense Gaussian and sub-gaussian sketches. We also show that some sparse sketches do not have this desired property. Then, we provide a non-trivial new construction and algorithm, using ideas from both data-oblivious projections and data-aware sampling, to get small inversion bias even for very sparse sketches.

1.1. Overview

Consider an $n \times d$ data matrix $\mathbf{A}$ of rank d, where $n \geq d$. In many applications, we wish to approximate quantities of the form $F((\mathbf{A}^\top \mathbf{A})^{-1})$, where $(\mathbf{A}^\top \mathbf{A})^{-1}$ is the $d \times d$ inverse data covariance and $F(\cdot)$ is a linear functional. Our goal is to provide a finer control over the effect of matrix inversion on the quality of such estimates. Here are some of the motivating examples:

- The vector $(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{b}$ is the solution of ordinary least squares (OLS) when $\mathbf{b} = \mathbf{A}^\top \mathbf{y}$ for a vector $\mathbf{y}$, arguably the most widely used multivariate statistical method (Anderson, 2003; Rao et al., 1973; Hastie et al., 2009), and it is also crucial for the Newton's method in numerical optimization (Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006). In particular, accurate approximations of this vector lead directly to improved convergence guarantees for many optimization algorithms (Pilanci and Wainwright, 2016; Wang et al., 2018b).

- The scalar $\mathbf{x}^\top (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{x}$ for a vector $\mathbf{x}$, has numerous use-cases: When $\mathbf{x} = \mathbf{a}_i$ is one of the rows of $\mathbf{A}$, then it represents the statistical leverage scores (Drineas et al., 2006b); If $\mathbf{x} = \mathbf{e}_i$ is a standard basis vector, then this is the squared length of the confidence interval for the i-th coefficient in OLS (Anderson, 2003; Hastie et al., 2009).

- The scalar $\operatorname{tr} \mathbf{C}(\mathbf{A}^\top \mathbf{A})^{-1}$ for a matrix $\mathbf{C}$, is used to quantify uncertainty in statistical results, e.g., via the mean squared error (MSE) of estimating the regression coefficients in OLS (Anderson, 2003; Hastie et al., 2009), and to formulate widely used criteria from experimental design, e.g., A-designs and V-designs (Pukelsheim, 2006; Cox and Reid, 2000).

More generally, our work is also motivated by the important problem of inverse covariance estimation in statistics, machine learning, finance, signal processing, and related areas (Dempster, 1972; Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Lam and Fan, 2009; Cai et al., 2011;

Ledoit and Wolf, 2012; Marjanovic and Hero, 2015). In this area, we wish to estimate statistically the inverse covariance matrix of a population, or some of its functionals, based on a finite number of samples. Furthermore, inverting covariance matrices occurs in Bayesian statistics (Hartigan, 1969; Gelman et al., 2013), Gaussian processes (Rasmussen, 2003), as well as time series analysis and control, e.g., via the Kalman filter (Welch and Bishop, 1995; Brockwell and Davis, 2009).

When n and d are large, and particularly when $n \gg d$, then the costs of storing the matrix $\mathbf{A}$ and of computing $(\mathbf{A}^\top\mathbf{A})^{-1}$ are prohibitively large. Matrix sketching has proven successful at drastically reducing these costs by approximating the inverse covariance with a sketched estimate $(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$ based on a smaller matrix $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$, where $\mathbf{S}$ is a random $m \times n$ matrix and $m \ll n$ (Mahoney, 2011; Halko et al., 2011b; Woodruff, 2014; Drineas and Mahoney, 2016, 2018). As a concrete algorithmic motivation for our work, consider the following popular strategy for boosting the quality of such estimates: Construct multiple copies in parallel, based on independent sketches, and then average the estimates. This strategy is especially useful in distributed architectures, where storage and computing resources are spread out across many machines, and has commonly appeared in the literature (Konecný et al., 2016a,b; Wang et al., 2018b; Dereziński and Mahoney, 2019). While promising in practice, this averaging technique is fundamentally limited by the *inversion bias*: even though the sketched covariance estimate is unbiased, $\mathbb{E}[\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}] = \mathbf{A}^\top\mathbf{A}$, its inverse in general is not unbiased, i.e., $\mathbb{E}[(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}] \neq (\mathbf{A}^\top\mathbf{A})^{-1}$. When the sketch size m is not much larger than the dimension d, the size of this bias can be very significant, even as large as the approximation error, in which case averaging becomes ineffective. Motivated by this, we ask:

When is the inversion bias small, relative to the approximation error?

In this paper, we develop a framework for analyzing the inversion bias of sketching, via the notion of an (ϵ, δ)-unbiased estimator (Definition 3), and we show how it can be used to provide improved approximation guarantees for averaging. Through this framework, we provide several contributions towards addressing the above question.

Sub-gaussian sketches have small inversion bias. Arguably the most classical family of sketches consists of dense random matrices $\mathbf{S}$ with i.i.d. sub-gaussian entries. These sketches offer strong relative error approximation guarantees via the so-called *subspace embedding* property, at the expense of high computational cost of the matrix product $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$. We show that, upon a simple correction, sub-gaussian sketches are nearly-unbiased, i.e., their inversion bias is much smaller than the approximation error, which means that averaging can be used to significantly improve the approximation quality. In particular, we show that, after a simple scalar rescaling, the inverse covariance estimator of the form $(\frac{m}{m-d}\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$ achieves ϵ inversion bias relative to $(\mathbf{A}^\top\mathbf{A})^{-1}$ with a sketch of size only $m = O(d + \sqrt{d}/\epsilon)$ (Proposition 4). In contrast, to ensure that $(\frac{m}{m-d}\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$ is an η relative error approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$ via the subspace embedding property, we need a sub-gaussian sketch of size $m = \Theta(d/\eta^2)$, which is comparatively larger if we let $\eta = \Theta(\epsilon)$. This implies that an aggregate estimator obtained via averaging can with high probability produce a relative error approximation that is by a factor of $O(1/\sqrt{m})$ better than the approximation error offered by any one of the estimators being averaged.

LEverage Score Sparsified (LESS) embeddings. We show that existing algorithmically efficient sketching techniques may not provide guarantees for the inversion bias that match those satisfied by dense sub-gaussian sketches (see Theorem 10 for a lower bound on leverage score sampling, and a discussion of other methods in Appendix C.3). To address this, we propose a new family of sketching methods, called LEverage Score Sparsified (LESS) embeddings, which combines a data-oblivious sparsification strategy reminiscent of the CountSketch with the data-aware approach of approximate leverage score sampling. LESS embeddings have time complexity $O(\mathrm{nnz}(\mathbf{A})\log n + md^2)$ and achieve ϵ inversion bias with the sketch of size $m = O(d\log d + \sqrt{d}/\epsilon)$, nearly matching our guarantee for sub-gaussian sketches (Theorem 8). Thus, our new algorithm provides a promising way to address the fundamental problem of inversion bias, and it may have many other applications in the future. Finally, our analysis reveals two structural conditions for small inversion bias (Theorem 11), one of which (Condition 2, called the Restricted Bai-Silverstein condition) leads to a generalization of a classical inequality used in random matrix theory, and should be of independent interest.

1.2. Related work

Distributed averaging. Averaging strategies have been studied extensively in the literature, particularly in the context of machine learning and numerical optimization. This line of work has proven particularly effective for *federated learning* (Konecný et al., 2016a,b), where local storage and communication bandwidth are particularly constrained. The performance of averaged estimates was analyzed in numerous statistical learning settings (McDonald et al., 2009, 2010; Zhang et al., 2013; Dobriban and Sheng, 2018, 2020) and in stochastic first-order optimization (Zinkevich et al., 2010; Agarwal and Duchi, 2011). Of particular relevance to our results is a recent line of works on distributed second-order optimization (Shamir et al., 2014; Zhang and Lin, 2015; Reddi et al., 2016; Wang et al., 2018b), as well as large-scale second-order optimization (Yao et al., 2019, 2020), since sketching is used there to estimate (implicitly) the inverse Hessian matrix which arises in Newton-type methods. In particular, Dereziński and Mahoney (2019); Dereziński et al. (2020a) pointed to Hessian inversion bias as a key challenge in these approaches. To address it, their algorithms use non-i.i.d. sampling sketches based on Determinantal Point Processes (DPPs) (Dereziński and Mahoney, 2021). DPP-based sketches are known to correct inversion bias exactly (Dereziński and Warmuth, 2018; Dereziński et al., 2019c,a). However, state-of-the-art DPP sampling algorithms (Dereziński et al., 2018, 2019; Calandriello et al., 2020) have time complexity $O(\mathrm{nnz}(\mathbf{A})\log n + d^4 \log d)$, which is considerably more expensive than fast sketching techniques when dimension d is large.

Random matrix theory. When considering $\mathbf{S} \in \mathbb{R}^{m\times n}$ having i.i.d. zero-mean rows, $\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}$ can be viewed as the popular *sample covariance estimator* of the "population covariance matrix" $\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{d\times d}$. In this area, one often considers the matrix $(\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A} - z\mathbf{I})^{-1}$ for $z \in \mathbb{C}\setminus\mathbb{R}_+$, the so-called *resolvent* matrix, which plays a fundamental role in the literature of random matrix theory (RMT) (Marchenko and Pastur, 1967; Bai and Silverstein, 2010; Edelman and Rao, 2005; Anderson et al., 2010; Couillet and Debbah, 2011; Tao, 2012; Bun et al., 2017) and which is directly connected to the popular Marchenko-Pastur law (Marchenko and Pastur, 1967). The RMT literature focuses on the Stieltjes transform (that is, the normalized trace of the resolvent) to investigate the limiting eigenvalue distribution of large random matrices of the form $\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}$ as $m, n, d \to \infty$ at the same rate. Here, we provide *precise* and *finite-dimensional* results on the inverse sketched matrix. This addresses the important case of $z = 0$, which is typically avoided in RMT analyses, due to the difficulty of dealing with the possible singularity. More generally, the resolvent also appears as the key object of study in the spectrum analysis of linear operators in general Hilbert space (Akhiezer and Glazman, 2013), as well as in modern convex optimization theory (Bauschke and Combettes, 2011), thereby showing a much broader interest of the proposed analysis.

Sketching. For overviews of sketching and random projection methods, we refer to (Vempala, 2005; Halko et al., 2011b; Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016, 2017, 2018; Dereziński and Mahoney, 2021). A key result in this area is the Johnson-Lindenstrauss lemma, which states that norms, and thus also relative distances between points, are *approximately* preserved after sketching, i.e., $(1-\eta)\|\mathbf{x}_i\|^2 \le \|\mathbf{S}\mathbf{x}_i\|^2 \le (1+\eta)\|\mathbf{x}_i\|^2$ for $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^p$. This is further extended to the *subspace embedding property*: for all $\mathbf{x}$, the norm of $\mathbf{x}$ is preserved up to an η factor. Subspace embeddings were first used in RandNLA by Drineas et al. (2006b), where they were used in a data-aware context to obtain relative-error approximations for ℓ_2 regression and low-rank matrix approximation (Drineas et al., 2008). Subsequently, data-oblivious subspace embeddings were used by Sarlos (2006) and popularized by Woodruff (2014). Both data-aware and data-oblivious subspace embeddings can be used to derive bounds for the accuracy of various algorithms (Drineas and Mahoney, 2016, 2018).

The most popular sketching methods include random projections with i.i.d. entries, random sampling of the datapoints, uniform orthogonal projections, Subsampled Randomized Hadamard Transform (SRHT) (Sarlos, 2006; Ailon and Chazelle, 2006), leverage score sampling (Drineas et al., 2008, 2012; Ma et al., 2015), and CountSketch (Charikar et al., 2002; Clarkson and Woodruff, 2017; Nelson and Nguyên, 2013; Meng and Mahoney, 2013). Random projection based approaches have been developed for a wide variety of problems in data science, statistics, machine learning etc., including linear regression (Sarlos, 2006; Drineas et al., 2011; Raskutti and Mahoney, 2016; Dobriban and Liu, 2018), ridge regression (Lu et al., 2013; Chen et al., 2015; Wang et al., 2018a; Liu and Dobriban, 2019), two sample testing (Lopes et al., 2011; Srivastava et al., 2016), classification (Cannings and Samworth, 2017), PCA (Frieze et al., 2004; Drineas et al., 2006a; Sarlos, 2006; Liberty et al., 2007; Halko et al., 2011a,b; Woolfe et al., 2008; Musco and Musco, 2015; Tropp et al., 2017; Dasarathy et al., 2015; Yang et al., 2020; Gataric et al., 2020), convex optimization (Pilanci and Wainwright, 2015, 2016, 2017), etc.; see (Woodruff, 2014; Drineas and Mahoney, 2016, 2018) for a more comprehensive list. Our new LESS embeddings have the potential to be relevant for all those important applications.

2. Dense Gaussian and sub-gaussian sketches have small inversion bias

Consider first the classical Gaussian sketch, i.e., where the entries of $\mathbf{S}$ are i.i.d. standard normal scaled by $1/\sqrt{m}$. In this special case, the sketched covariance matrix $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is a Wishart-distributed random matrix, and we have:

$$\mathbb{E}\big[(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}\big] = \tfrac{m}{m-d-1}(\mathbf{A}^\top\mathbf{A})^{-1} \quad \text{for} \quad m \ge d+2. \tag{1}$$

In other words, even though the sketched inverse covariance is not an unbiased estimate, the bias can be corrected by simply scaling the matrix, after which averaging can be used effectively without encountering any inversion bias.

The key property which enables exact bias-correction for the Gaussian sketch is *orthogonal invariance*. This property requires that for any orthonormal matrix $\mathbf{O}$, the distributions of the random matrices $\mathbf{S}$ and $\mathbf{SO}$ are identical. An example beyond Gaussians are Haar sketches, which are uniform over all partial orthogonal matrices. If a sketch $\mathbf{S}$ is orthogonally invariant and $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is invertible with probability one, then we can show that the inversion bias can be corrected exactly, in that, (1) holds with some constant factor c (replacing the factor $\frac{m}{m-d-1}$) that depends on the distribution of the sketch (see Proposition 38 in Appendix G).

Exact bias-correction, achieved by the Gaussian sketch and other orthogonally invariant sketches, is no longer possible for general sub-gaussian sketches. Here, we consider sketching matrices with i.i.d. entries that (after scaling by $\sqrt{m}$) have $O(1)$ sub-gaussian Orlicz norm. Consider for example the so-called Rademacher sketch, with $\mathbf{S}$ consisting of scaled i.i.d. random sign entries (which is useful for reducing the cost of randomness relative to the Gaussian sketch). In this case, an exact bias-correction analogous to (1) is clearly infeasible for any $d > 1$, simply because, with some positive (but exponentially small) probability, the matrix $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ will be non-invertible, making the expectation undefined. Yet, any task where we observe at most polynomially many independent estimates (such as averaging) should not be affected by such low-probability events, so we need a notion of near-unbiasedness that is robust to this. To that end, we first recall a standard definition of a relative error approximation for a positive semi-definite matrix.

Definition 1 (Relative error approximation) *A positive semi-definite (p.s.d.) matrix $\tilde{\mathbf{C}}$ (or a non-negative scalar) is an η-approximation of $\mathbf{C}$, denoted as $\tilde{\mathbf{C}} \approx_\eta \mathbf{C}$, if*

$$\mathbf{C}/(1+\eta) \preceq \tilde{\mathbf{C}} \preceq (1+\eta)\cdot\mathbf{C}.$$

If $\tilde{\mathbf{C}}$ is random and the above holds with probability $1-\delta$, then we call it an (η,δ)-approximation.

Remark 2 (Subspace embedding) *If $\tilde{\mathbf{C}} = \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ where $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times d}$ is a sketch of $\mathbf{A} \in \mathbb{R}^{n\times d}$, then the condition $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} \approx_\eta \mathbf{A}^\top\mathbf{A}$ is called the subspace embedding property with error η.*

For instance, any sketching matrix $\mathbf{S}$ with i.i.d. $O(1)$ sub-gaussian random entries, of size $m = O((d+\ln(1/\delta))/\eta^2)$, where $\eta \in (0,1)$, ensures that $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$ with probability $1-\delta$ satisfies the subspace embedding property with error η. In other words, $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is an (η,δ)-approximation of $\mathbf{A}^\top\mathbf{A}$ (This is known to be tight; see, e.g., Nelson and Nguyen, 2014). As a consequence, the same guarantee applies to the inverse $(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$, relative to $(\mathbf{A}^\top\mathbf{A})^{-1}$. The δ failure probability makes this definition robust to the rare events where $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is not invertible. It is natural to desire a similar robustness in the definition of near-unbiasedness. We achieve this as follows.

Definition 3 ((ϵ,δ)-unbiased estimator) *A random p.s.d. matrix $\tilde{\mathbf{C}}$ is an (ϵ,δ)-unbiased estimator of $\mathbf{C}$ if there is an event $\mathcal{E}$ that holds with probability $1-\delta$ such that*

$$\mathbb{E}\big[\tilde{\mathbf{C}} \mid \mathcal{E}\big] \approx_\epsilon \mathbf{C}, \quad \textit{and} \quad \tilde{\mathbf{C}} \preceq O(1)\cdot\mathbf{C} \quad \textit{when conditioned on } \mathcal{E}.$$

Note that this definition only becomes meaningful if we use it with an ϵ that is much smaller than the approximation error η in Definition 1 (for instance, we will often have $\eta = \Omega(1)$ and $\epsilon \ll 1$). Further, note the following two important aspects of Definition 3. First, instead of a simple expectation, we condition on some high probability event $\mathcal{E}$, which, similarly as in Definition 1, allows robustness against such corner cases as when the sketch $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$ is not invertible. Second, conditioned on the event $\mathcal{E}$, in addition to an ϵ-approximation holding in expectation, we require a weaker upper bound to hold almost surely, in terms of the target matrix $\mathbf{C}$ scaled by some constant factor. This condition is important to guard against certain corner cases where the probability mass is extremely skewed. For instance, suppose that $\tilde{C}$ is a scalar random variable which is uniform over $[0,1]$ and has an additional probability mass of 10^{-10} at the value 10^{100}. Here, averaging will not prove effective at converging to the true expectation of $\tilde{C}$, but we could still use the notion of (ϵ,δ)-unbiasedness to show that the average of an appropriately chosen number (much smaller than

10^{10}) of i.i.d. copies will converge very close to 0.5, by choosing an event $\mathcal{E}$ that avoids the 10^{100} (see Appendix E).

We are now ready to state our main result for sub-gaussian sketches (this is in fact a corollary of our more general result, Theorem 11, discussed in Section 4), which asserts that after proper rescaling, not only the Gaussian sketch, but in fact all sub-gaussian sketches (including the Rademacher sketch) enjoy small inversion bias.

Proposition 4 (Near-unbiasedness of sub-gaussian sketches) *Let $\mathbf{S}$ be an $m \times n$ random matrix such that $\sqrt{m}\,\mathbf{S}$ has i.i.d. $O(1)$-sub-gaussian entries with mean zero and unit variance. There is $C = O(1)$ such that for any $\epsilon, \delta \in (0,1)$ if $m \geq C(d + \sqrt{d}/\epsilon + \log(1/\delta))$, then for all $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d, $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an (ϵ,δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$.*

Observe that the scaling $\frac{m}{m-d}$ essentially matches the exact bias-correction for Gaussian sketches, which is $\frac{m}{m-d-1}$. In fact, the same statement of the theorem holds with either scaling, and we merely chose the simplest form of the scaling.

As a corollary of the near-unbiasedness of sub-gaussian sketches, we can show the following approximation guarantee for averaging the inverse covariance matrix estimates. Recall that our primary motivation is parallel and distributed averaging, where the computational cost does not grow with the number of independent estimates.

Corollary 5 *Let $\mathbf{S}$ be a sub-gaussian sketching matrix of size m, and let $\mathbf{S}_1, ..., \mathbf{S}_q$ be i.i.d. copies of $\mathbf{S}$. There is $C = O(1)$ such that if $m \geq C(d + \sqrt{d}/\epsilon + \log(q/\delta))$ and $q \geq Cm\log(d/\delta)$, then for any $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d, $\frac{1}{q}\sum_{i=1}^q(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_i^\top\mathbf{S}_i\mathbf{A})^{-1}$ is an (ϵ,δ)-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$.*

Proposition 4 shows that for a sub-gaussian sketch $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$ of size $m \geq Cd$, the sketched inverse covariance $(\frac{m}{m-d}\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$ has inversion bias $O(\sqrt{d}/m)$. This means that the inversion bias of this estimator is smaller than the approximation error, which is $\Theta(\sqrt{d/m})$, by a factor of $O(1/\sqrt{m})$. Thus, using Corollary 5, we can reduce the approximation error by averaging $q = O(m\log(d/\delta))$ copies of this estimator, obtaining that $\frac{1}{q}\sum_{i=1}^q(\frac{m}{m-d}\tilde{\mathbf{A}}_i^\top\tilde{\mathbf{A}}_i)^{-1}$ is with high probability an $O(\sqrt{d}/m)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$. In particular, when $m = \Theta(d)$, then the approximation error of a single estimate (without averaging) is $\Theta(1)$, whereas the approximation error of the averaged estimate is only $O(1/\sqrt{d})$.

3. Main results: Less inversion bias with LESS embeddings

To address the high computational cost of sub-gaussian sketches, while preserving their good near-unbiasedness properties, we propose a new family of sketches, which we call LEverage Score Sparsified (LESS) embeddings. A LESS embedding is defined simply as a sparsified sub-gaussian sketch, where the sparsification is designed so as to ensure small inversion bias for a particular matrix $\mathbf{A}$. Our approach combines ideas from approximate leverage score sampling (which is data-aware) with ideas from sparse embedding matrices (which are normally data-oblivious). Importantly, neither strategy by itself is sufficient to ensure small inversion bias (see our lower bound in Theorem 10 and discussion in Section 4). Each row of a LESS embedding is sparsified independently using a sparsification pattern defined as follows. Recall that for a tall full rank matrix $\mathbf{A}$, we use $\mathbf{a}_i^\top$ to denote the ith row of $\mathbf{A}$, and the ith leverage score of $\mathbf{A}$ is defined as $l_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_i$.

Definition 6 (LESS: LEverage Score Sparsified embedding) *Fix a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d with leverage scores $l_1, ..., l_n$, and let $s_1, ..., s_d$ be sampled i.i.d. from a probability distribution $(p_1, ..., p_n)$ such that $p_i \approx_{O(1)} l_i/d$ for all i. Then, the random vector $\boldsymbol{\xi}^\top = \big(\sqrt{\frac{b_1}{dp_1}}, ..., \sqrt{\frac{b_n}{dp_n}}\big)$, where $b_i = \sum_{t=1}^d 1_{[s_t=i]}$, is called a <u>leverage score sparsifier</u> for $\mathbf{A}$.*

Sketching matrix $\mathbf{S}$ is a <u>LESS embedding</u> of size m for a matrix $\mathbf{A}$, if it consists of m i.i.d. row vectors distributed as $\frac{1}{\sqrt{m}}(\mathbf{x}\circ\boldsymbol{\xi})^\top$, where $\circ$ denotes an entry-wise product and $\mathbf{x}$ is a random vector with i.i.d. mean zero, unit variance, $O(1)$-sub-gaussian entries.

Remark 7 (Time complexity of LESS) *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d, there is an algorithm with an $O(\mathrm{nnz}(\mathbf{A})\log n + d^3 \log d)$ time preprocessing step, that can then construct a LESS embedding sketch $\mathbf{SA}$ of size m in time $O(md^2)$. In the following results we always use $m \geq d\log d$, in which case the total cost of constructing a LESS embedding is $O(\mathrm{nnz}(\mathbf{A})\log n + md^2)$.*

The matrix product $\mathbf{SA}$ costs only $O(md^2)$ because, by definition, the number of non-zeros per row of $\mathbf{S}$ is bounded almost surely by d. It is not essential for our analysis that we sample exactly d indices in each row of a LESS embedding, but we fix it here for the sake of simplicity. We could also have approximately d non-zeros per row, and similar results would still hold. To construct the distribution $(p_1, ..., p_n)$, the sparsifier requires a constant relative error approximation of all the leverage scores of $\mathbf{A}$, which can be computed in $O(\mathrm{nnz}(\mathbf{A})\log n + d^3 \log d)$ time (Drineas et al., 2012; Clarkson and Woodruff, 2017). Alternatively we can use our approach in a data-oblivious way, by combining LEverage Score Sparsification with the Randomized Hadamard Transform (Ailon and Chazelle, 2009; Drineas et al., 2011), which we may abbreviate as LESS-RHT. Here, the matrix $\mathbf{A}$ is first transformed so that it has approximately uniform leverage scores (Drineas et al., 2012), and then we can sparsify it using a uniform distribution, i.e., $p_i = 1/n$ for all i, with total cost $O(nd\log n + md^2)$. Finally, computing the sketched inverse covariance matrix estimator $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{SA})^{-1}$ only adds an $O(md^2)$ cost. These costs can be further optimized using fast matrix multiplication (Williams, 2012).[1]

In our main result, we show that LESS embeddings enjoy small inversion bias, nearly matching our guarantee for sub-gaussian sketches (Proposition 4).

Theorem 8 (Near-unbiasedness for LESS) *Suppose that $\mathbf{S}$ is a LESS embedding of size m for a rank d matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. There is $C = O(1)$ such that if $m \geq C(d\log(d/\delta) + \sqrt{d}/\epsilon)$ then the sketch $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{SA})^{-1}$ is an (ϵ, δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$.*

Thus, we show that the inversion bias guarantee for LESS embeddings matches our result for sub-gaussian sketches up to a logarithmic factor. This additional factor is standard in the analysis of fast sketching methods. It comes from the fact that, as an artifact of the matrix concentration bounds (Tropp, 2012) we use in our analysis of LESS embeddings, a sketch of size $m = O(d\log d)$ is needed to satisfy the subspace embedding property, which is one of our two structural conditions for small inversion bias (see Section 4). As a corollary, we obtain an improved guarantee for parallel and distributed averaging of i.i.d. sketched inverse covariance estimates which also matches the corresponding statement for sub-gaussian sketches (Corollary 5) up to logarithmic factors.

[1] The cost of computing the matrix product $\mathbf{SA}$ can be optimized beyond $O(md^2)$ by adapting the fast matrix multiplication routines to take advantage of the sparsity pattern; see, e.g., Yuster and Zwick (2005).

Corollary 9 *Let $\mathbf{S}$ be a LESS embedding matrix of size m for a rank d matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, and let $\mathbf{S}_1, ..., \mathbf{S}_q$ be i.i.d. copies of $\mathbf{S}$. There is $C = O(1)$ such that if $m \geq C(d\log(q/\delta) + \sqrt{d}/\epsilon)$ and $q \geq Cm\log^2(d/\delta)$, then $\frac{1}{q}\sum_{i=1}^q(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_i^\top\mathbf{S}_i\mathbf{A})^{-1}$ is an (ϵ, δ)-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$.*

To motivate and place our new algorithm into context, we demonstrate that existing fast sketching techniques may not achieve an inversion bias bound comparable to that of sub-gaussian sketches, even if they achieve a nearly matching subspace embedding guarantee. This lower bound demonstrates the hardness of constructing an (ϵ, δ)-unbiased estimator of the inverse covariance matrix from its sketch. We show this here for leverage score sampling (Drineas et al., 2006b, 2008, 2012; Ma et al., 2015). However, based on evidence from our analysis, we conjecture that similar lower bounds hold for other methods such as Subsampled Randomized Hadamard Transform (Ailon and Chazelle, 2009; Drineas et al., 2011) and data-oblivious sparse embedding matrices (Clarkson and Woodruff, 2017; Nelson and Nguyên, 2013; Meng and Mahoney, 2013).[2]

Theorem 10 (Lower bound for leverage score sampling) *For any $n \geq 2d \geq 4$, there is an $n \times d$ matrix $\mathbf{A}$ and a row sampling $(p_1, ..., p_n)$, with a corresponding $m \times n$ sketching matrix $\mathbf{S}$, s.t.:*

1. *The row sampling $(p_1, ..., p_n)$ is a $1/2$-approximation of leverage score sampling; and*
2. *For any sketch size m and scaling γ, $(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is not an (ϵ, δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ with any $\epsilon \leq c\frac{d}{m}$ and $\delta \leq c(\frac{d}{m})^2$, where $c > 0$ is an absolute constant.*

In the proof of Theorem 10, we develop a new lower bound for the inverse moment of the Binomial distribution (Lemma 36), by using anti-concentration of measure via the Paley-Zygmund inequality, which should be of independent interest. To illustrate Theorem 10, consider a sketch of size $m = O(d\log d)$. This is sufficient to ensure that approximate leverage score sampling achieves the subspace embedding property with relative error $O(1)$. In particular, it implies that for any $\gamma = \Theta(1)$, the inverse covariance matrix estimator $(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is with high probability an $O(1)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$. Our lower bound implies that the inversion bias of any such estimator is $\Omega(1/\log d)$, which is up to logarithmic factors the same as the approximation achieved by a single estimator.

Thus, Theorem 10 shows that when $m = O(d\log d)$, averaging i.i.d. copies of the sketched inverse covariance estimator obtained from approximate leverage score sampling may lead to only $\Omega(1/\log d)$ factor improvement in the approximation, which is merely inverse-logarithmic in d. In contrast, Theorem 8 shows that, when using our new LESS embeddings with the same sketch size and time complexity, averaging i.i.d. copies of the sketched inverse covariance reduces the approximation error by a factor of $O(1/\sqrt{d})$, which is inverse-polynomial in d and thus far superior to what is achievable by approximate leverage score sampling.

4. Our techniques: Structural conditions for near-unbiasedness

In order for our analysis of inversion bias to apply to a wide range of sketching techniques, we give two key structural conditions for a random sketching matrix $\mathbf{S}$ that are sufficient to achieve provably

[2] An alternative approach to achieving small inversion bias is to chain together a fast sketch having a larger size, say, $t = \widetilde{O}(d/\epsilon^2)$, with a sub-gaussian sketch having a smaller size $m = O(d + \sqrt{d}/\epsilon)$. However, this leads to a suboptimal time complexity in terms of the polynomial dependence on d due to the cost $O(tmd)$ of the sub-gaussian sketch. For example, with $\epsilon = 1/\sqrt{d}$, the overall cost is $\widetilde{O}(\mathrm{nnz}(\mathbf{A}) + d^4)$ compared to $\widetilde{O}(\mathrm{nnz}(\mathbf{A}) + d^3)$ with LESS.

small inversion bias. The first is the subspace embedding property discussed in Remark 2, which we now use as one of the key conditions needed in our analysis.

Condition 1 (Subspace embedding) *The (sketching) matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ satisfies the subspace embedding condition with $\eta \geq 0$ for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, if $\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A} \approx_\eta \mathbf{A}^\top\mathbf{A}$.*

The second structural condition for small inversion bias is a property of each individual row of $\mathbf{S}$. We use an n-dimensional random row vector $\mathbf{x}^\top$ to denote the marginal distribution of a row of $\mathbf{S}$ (after scaling by $\sqrt{m}$). This condition represents a key novelty in our analysis.

Condition 2 (Restricted Bai-Silverstein) *The random vector $\mathbf{x} \in \mathbb{R}^n$ satisfies the Restricted Bai-Silverstein condition with $\alpha > 0$ for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, if $\mathrm{Var}\big[\mathbf{x}^\top\mathbf{B}\mathbf{x}\big] \leq \alpha \cdot \mathrm{tr}(\mathbf{B}^2)$ for all p.s.d. matrices $\mathbf{B}$ such that $\mathbf{B} = \mathbf{P}\mathbf{B}\mathbf{P}$, where $\mathbf{P}$ is the projection onto the column span of $\mathbf{A}$.*

Based on these two structural conditions, we show the following result, which we use to prove both Proposition 4 and Theorem 8. In this result, we will refer to an $m \times n$ sketching matrix $\mathbf{S}_m$, indexed by the number of rows m.

Theorem 11 (Structural conditions for near-unbiasedness) *Fix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with rank d and let $\mathbf{S}_m$ consist of $m \geq 8d$ i.i.d. rows distributed as $\frac{1}{\sqrt{m}}\mathbf{x}^\top$, where $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_n$. Suppose that $\mathbf{S}_{m/3}$ satisfies Condition 1 (subspace embedding) for $\eta = 1/2$, with probability $1-\delta/3$, where $\delta \leq 1/m^3$. Suppose also that $\mathbf{x}$ satisfies Condition 2 (Restricted Bai-Silverstein) with some $\alpha \geq 1$. Then $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A})^{-1}$ is an (ϵ,δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ for $\epsilon = O(\alpha\sqrt{d}/m)$.*

The proof of Theorem 11 adapts and extends techniques for analyzing the limiting Stieltjes transform for high-dimensional random matrices in the so-called Marchenko-Pastur regime (also called the proportional or mean-field limit). This regime arises if we let n, m and d all go to infinity and let the ratio m/d converge to a fixed constant larger than unity. Crucially, our analysis is non-asymptotic, and it is not restricted to the constant aspect ratio between the sketch size and the dimension. Further, while classical random matrix theory analysis considers matrix resolvents, which take the form $(\gamma\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A} + z\mathbf{I})^{-1}$ for $z, \gamma \neq 0$, and are well-defined with full probability, we consider the case of $z = 0$ where the matrix in question may be undefined with positive probability. We address this by defining a high probability event which ensures that the sketch $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A})^{-1}$ is well-defined and bounded, while preserving enough of the independence structure in the conditional distribution for the expectation analysis to go through. Specifically, we split the sketch into three parts, and we condition on the event that each part satisfies the subspace embedding property. This way, for any pair of rows, there is a part of the sketch that ensures invertibility while being independent from the two rows, which is important for the analysis.

Subspace embedding condition. Our first structural condition for small inversion bias (Condition 1) is a variant of the subspace embedding property, which is standard in the sketching literature. In particular, this condition immediately implies that $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A})^{-1}$ is with probability $1-\delta$ an $O(1)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$. For sub-gaussian sketches this is known to hold with sketch size $O(d + \log(1/\delta))$ (Nelson and Nguyen, 2014). We prove this for LESS of size $O(d\log(d/\delta))$.

Lemma 12 (Subspace embedding for LESS) *Suppose that $\mathbf{S}$ is a LESS embedding of size m for a rank d matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. There is $C = O(1)$ such that if $m \geq Cd\log(d/\delta)/\eta^2$ for $\eta \in (0,1)$, then the sketch $\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}$ is an (η,δ)-approximation of $\mathbf{A}^\top\mathbf{A}$.*

The subspace embedding guarantee for LESS embeddings is as good as that for existing fast sketching methods. However, the analysis differs from the ones used for either data-aware leverage score sampling or for data-oblivious sparse sketches. We show the result by deriving a subexponential bound on the matrix moments of a LESS embedding (Lemma 30), relying on a novel variant of the Hanson-Wright concentration inequality for quadratic forms based on orthogonal projection matrices (Lemma 31). We then use this to invoke a matrix Bernstein inequality for random matrices with subexponential moments (Tropp, 2012, Theorem 6.2).

Restricted Bai-Silverstein condition. Our second structural condition for small inversion bias (Condition 2) is not commonly seen in sketching, but we expect that it will be of broader interest in adapting high-dimensional random matrix theory to RandNLA (Dereziński et al., 2020b, 2019b; Dereziński and Mahoney, 2021). It is based on the classical inequality of Bai and Silverstein (Bai and Silverstein, 2010) which bounds the deviation of a random quadratic form $\mathbf{x}^\top\mathbf{B}\mathbf{x}$ from its mean. We call it the *Restricted Bai-Silverstein condition* because, unlike in the classical version, we only require the inequality to hold for matrices $\mathbf{B}$ that are restricted to the subspace spanned by the columns of $\mathbf{A}$. By contrast, in classical random matrix theory it is often assumed that the the following (unrestricted) condition holds.

Condition 3 (Bai-Silverstein) *Random vector $\mathbf{x}\in\mathbb{R}^n$ satisfies the (unrestricted) Bai-Silverstein condition with $\alpha>0$, if $\mathrm{Var}\big[\mathbf{x}^\top\mathbf{B}\mathbf{x}\big]\leq\alpha\cdot\mathrm{tr}(\mathbf{B}^2)$ for all $n\times n$ p.s.d. matrices $\mathbf{B}$.*

When the random vector $\mathbf{x}$ is $O(1)$-sub-gaussian, then Condition 3 is satisfied with $\alpha=O(1)$, as a consequence of the original inequality of Bai and Silverstein (2010).[3]

Lemma 13 (Bai-Silverstein inequality) *Let $\mathbf{x}$ have n independent entries with mean zero and unit variance such that $\mathbb{E}[x_i^4]=O(1)$. Then, Condition 3 is satisfied with $\alpha=O(1)$.*

5. Restricted Bai-Silverstein inequality

The Bai-Silverstein inequality from Lemma 13 does not directly apply to any of the fast sketching methods discussed above (see Appendix C.3 for lower bounds). However, we state and prove a generalization of this lemma, which allows us to show the Restricted Bai-Silverstein condition (Condition 2) for our new LESS embeddings.

To provide some intuition behind this result, consider the variance term $\mathrm{Var}[\mathbf{x}^\top\mathbf{B}\mathbf{x}]$ which appears in the Restricted Bai-Silverstein condition, where $\frac{1}{\sqrt{m}}\mathbf{x}^\top$ represents a random row vector of the sketching matrix $\mathbf{S}$. The condition requires that just this one row vector carries enough randomness to produce an accurate sketch of the trace of a quadratic form $\mathbf{B}$. This is in contrast to the subspace embedding condition, which uses the joint randomness of all the rows of $\mathbf{S}$. Lemma 13 achieves this by enforcing a fourth-moment bound on all of the entries of $\mathbf{x}$. Suppose that we sparsify this vector, following the strategy of sparse embedding matrices, by multiplying each entry of $\mathbf{x}$ with an independent scaled Bernoulli variable, obtaining $\sqrt{\frac{m}{s}}b_ix_i$ for $b_i\sim\mathrm{Bernoulli}(\frac{s}{m})$, where $s\ll m$ is the sparsity level and i is the entry index.[4] This preserves the mean and variance assumptions from Lemma 13, but as long as $s=o(m)$, it violates the fourth-moment assumption.

[3] The original lemma applies more broadly to higher moments; we cite only the case relevant to our analysis.

[4] Most commonly studied sparse embedding matrices have non-independent entries. However, the i.i.d. variant we consider offers an equivalent guarantee for the subspace embedding property. See Cohen (2016) and Appendix C.3.

Thus, it is natural to ask whether we can relax this fourth-moment assumption. It turns out that, if we do the sparsification in a data-oblivious manner, then the answer is no, since the random vector may not capture most of the relevant directions in the matrix $\mathbf{B}$ (see Appendix C.3). Importantly, this can occur even when the rows of the sketch *together* capture all of the directions, ensuring the subspace embedding property, which is already the case when we set the sparsity level to be as small as $s = O(\log d)$. In other words, there is a wide gap between the sparsity needed to preserve the Bai-Silverstein inequality, $s = \Omega(m)$, and sparsity needed to ensure the subspace embedding.

Crucially, Theorem 11 does not require the Bai-Silverstein inequality to hold for all $n \times n$ p.s.d. quadratic forms $\mathbf{B}$. Rather, it restricts the family of quadratic forms to those that lie within the column-span of the $n \times d$ data matrix $\mathbf{A}$. In particular, this restriction implies that the matrix $\mathbf{B}$ is low-rank (it has at most rank d) and its important directions are captured by the leverage scores of $\mathbf{A}$. We take advantage of this additional information to relax the fourth-moment assumptions, obtaining the following generalization of Lemma 13, which should be of independent interest.

Theorem 14 (Restricted Bai-Silverstein inequality) *Fix a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with rank d and leverage scores l_i, and let $\mathbf{x}$ have n independent entries with mean zero and unit variance such that $\mathbb{E}x_i^4 \leq C/l_i$. Then, $\mathbf{x}$ satisfies Condition 2 with $\alpha = C + 2$ for matrix $\mathbf{A}$.*

By setting $\mathbf{A} = \mathbf{I}_n$, where all leverage scores are 1 and the restriction on $\mathbf{B}$ is vacuous, we not only recover the statement of Lemma 13, but also our new analysis uses the Perron-Frobenius theorem to obtain a tight constant factor in the bound (see Appendix C). However, when $\mathbf{A}$ is a tall matrix, then the fourth-moment assumption becomes potentially much more broadly applicable (for example, when the leverage scores are uniform, we only need $\mathbb{E}x_i^4 \leq C \cdot n/d$). In particular, consider an i.i.d. sub-gaussian random vector $\mathbf{x}$ sparsified as follows: $\mathbf{x} \circ \boldsymbol{\xi}$, where we let $\xi_i = b_i/\sqrt{l_i}$ and $b_i \sim \mathrm{Bernoulli}(l_i)$. Then, the entries satisfy the assumptions of Theorem 14, with expected number of non-zeros equal to d. Note that this is different than the data-oblivious sparsification discussed above, since the entries of the vector corresponding to large leverage scores are less likely to be zeroed-out than others. This form of sparsification is nearly equivalent to the one we use for our LESS embeddings (see Definition 6; our analysis can be applied to either variant), except that it leads to a non-deterministic level of sparsification. In Appendix C.2 we prove the Restricted Bai-Silverstein condition with $\alpha = O(1)$ for a leverage score sparsified vector constructed as in Definition 6, which has non-independent entries.

6. Conclusions

We analyzed the phenomenon of inversion bias in sketching-based estimation tasks involving the inverse covariance matrix. Inversion bias is a significant bottleneck in methods that use parallel and distributed averaging. We showed that certain classical sketching methods (such as sub-gaussian sketches) have small inversion bias, while many algorithmically efficient sketches (such as leverage score sampling) may not provide such a guarantee. Finally, we developed a new efficient sketching method, called LEverage Score Sparsified (LESS) embeddings, which has small inversion bias and its computational cost is nearly-linear in the input size.

Estimation of the inverse covariance matrix and its various linear functionals is motivated by a rich body of literature in statistics, data science, numerical optimization, machine learning, signal processing, etc., which we summarized in detail in Section 1.1. Here, we additionally remark that the (ϵ, δ)-approximation guarantee we provide for the averaged estimates of the inverse covariance (see Corollaries 5 and 9) immediately implies corresponding approximation guarantees for linear functionals of the inverse covariance in numerous tasks. In a distributed environment, one can use this to build a system for querying such functionals, by aggregating coarse estimates computed locally from q sketches to produce an improved global estimate with minimal communication cost. We illustrate this here for a family of linear functionals of the form $\operatorname{tr}\mathbf{C}(\mathbf{A}^\top\mathbf{A})^{-1}$, parameterized by any p.s.d. matrix $\mathbf{C}$, as motivated by applications in statistical inference (see Section 1.1). The claim follows from Corollary 9 by letting $\mathbf{Q}_i = (\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_i^\top\mathbf{S}_i\mathbf{A})^{-1}$.

Corollary 15 (Querying linear functionals) *For any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\epsilon, \delta \in (0,1)$, we can use LESS embeddings of size $m = O(d\log(d/\epsilon\delta) + \sqrt{d}/\epsilon)$ to construct $\mathbf{Q}_1, ..., \mathbf{Q}_q \in \mathbb{R}^{d\times d}$ in parallel time $O(\operatorname{nnz}(\mathbf{A})\log n + md^2)$, where $q = O(m\log^2(d/\delta))$, so that with probability $1-\delta$:*

$$\text{For all p.s.d. matrices } \mathbf{C} \in \mathbb{R}^{d\times d}, \qquad \frac{1}{q}\sum_{i=1}^{q}\operatorname{tr}\mathbf{C}\mathbf{Q}_i \approx_\epsilon \operatorname{tr}\mathbf{C}(\mathbf{A}^\top\mathbf{A})^{-1}.$$

In the context of distributed optimization, our results can be directly applied to show improved convergence guarantees, for instance, in the case of the Distributed Iterative Hessian Sketch algorithm (Pilanci and Wainwright, 2016; Dereziński et al., 2020a) and Distributed Newton Sketch method (Wang et al., 2018b; Dereziński and Mahoney, 2019). Here, the quantity of interest is of the form $(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{b}$ for some vector $\mathbf{b}$ (where $\mathbf{A}^\top\mathbf{A}$ corresponds to the Hessian and $\mathbf{b}$ corresponds to the gradient). For those methods, an ϵ-approximation guarantee for the average of the sketched inverse covariance matrices, as in Corollaries 5 and 9, directly implies that the iterates $\mathbf{x}_t$ produced by the algorithms achieve a convergence rate of the form $\Delta_t \leq O(\epsilon^t)\cdot\Delta_0$, where Δ_t represents distance from the optimum in the t-th iteration. We illustrate this by applying Corollary 9 to the existing analysis of Distributed Newton Sketch, as outlined in Section 4 of Dereziński et al. (2020a), obtaining an improved linear-quadratic convergence rate for distributed empirical risk minimization.

Corollary 16 (Distributed Newton Sketch) *Consider a twice differentiable convex function $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbf{x}^\top\phi_i) + \frac{\lambda}{2}\|\mathbf{x}\|^2$, where $\mathbf{x}\in\mathbb{R}^d$ and $\phi_i^\top$ is the ith row of an $n\times d$ data matrix Φ. Given $\mathbf{x}_t$, we can use LESS embeddings to construct q independent randomized estimates $\widehat{\mathbf{H}}_1(\mathbf{x}_t), ..., \widehat{\mathbf{H}}_q(\mathbf{x}_t)$ of the Hessian $\nabla^2 f(\mathbf{x}_t)$ in parallel time $O(\operatorname{nnz}(\Phi)\log n + md^2)$, where $m = O(d\log(d/\epsilon\delta) + \sqrt{d}/\epsilon)$ and $q = O(m\log^2(d/\delta))$, so that*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{q}\sum_{i=1}^{q}\widehat{\mathbf{H}}_i(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t) \quad \text{with probability } 1-\delta \text{ satisfies:}$$

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\| \leq \max\left\{\epsilon\cdot\sqrt{\kappa}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|,\ \tfrac{2L}{\lambda_{\min}}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\right\} \quad \text{for} \quad \mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}),$$

where κ, L, $\lambda_{\min}$ are the condition number, Lipschitz constant and smallest eigenvalue of $\nabla^2 f(\mathbf{x})$.

This result provides an improvement over the recently proposed DPP-based sketching methods of Dereziński et al. (2020a), which suffer no inversion bias but are more expensive, as well as over other fast sketching methods like row sampling (e.g., see Wang et al., 2018b), which, as shown in this work, may indeed suffer from large inversion bias.

Acknowledgments

We would like to acknowledge DARPA, IARPA, NSF, and ONR via its BRC on RandNLA for providing partial support of this work. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 873–881. Curran Associates, Inc., 2011.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Naum Ilich Akhiezer and Izrail Markovich Glazman. *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Number 118. Cambridge University Press, 2010.

Theodore W Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer, 2009.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Donald L Burkholder. Distribution function inequalities for martingales. *the Annals of Probability*, pages 19–42, 1973.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Daniele Calandriello, Michał Dereziński, and Michal Valko. Sampling from a k-dpp without looking at all items. *arXiv preprint arXiv:2006.16947*, 2020. Accepted for publication, Proc. NeurIPS 2020.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

Shouyuan Chen, Yang Liu, Michael R Lyu, Irwin King, and Shengyu Zhang. Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210, 2015.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6):54:1–54:45, January 2017.

Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 278–287. SIAM, 2016.

Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

David Roxbee Cox and Nancy Reid. *The theory of the design of experiments*. CRC Press, 2000.

G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

Michał Dereziński and Michael W Mahoney. Distributed estimation of the inverse Hessian by determinantal averaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11401–11411. Curran Associates, Inc., 2019.

Michał Dereziński and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021.

Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2510–2519. Curran Associates, Inc., 2018.

Michał Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11542–11554. Curran Associates, Inc., 2019.

Michał Dereziński, Kenneth L. Clarkson, Michael W. Mahoney, and Manfred K. Warmuth. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1050–1069, Phoenix, USA, 25–28 Jun 2019a.

Michał Dereziński, Feynman Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019b. Accepted for publication, Proc. NeurIPS 2020.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Unbiased estimators for random design regression. *arXiv e-prints*, art. arXiv:1907.03411, Jul 2019c.

Michał Dereziński, Burak Bartan, Mert Pilanci, and Michael W Mahoney. Debiasing distributed second order optimization with surrogate sketching and scaled regularization. *arXiv preprint arXiv:2007.01327*, 2020a. Accepted for publication, Proc. NeurIPS 2020.

Michał Dereziński, Feynman Liang, Zhenyu Liao, and Michael W Mahoney. Precise expressions for random projections: Low-rank approximation and randomized Newton. *arXiv preprint arXiv:2006.10653*, 2020b. Accepted for publication, Proc. NeurIPS 2020.

Edgar Dobriban and Sifan Liu. Asymptotic for sketching in least squares regression. *arXiv preprint arXiv:1810.06089, NeurIPS 2019*, 2018.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412, to appear in The Annals of Statistics*, 2018.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

P. Drineas and M. W. Mahoney. Lectures on randomized numerical linear algebra. In M. W. Mahoney, J. C. Duchi, and A. C. Gilbert, editors, *The Mathematics of Data*, IAS/Park City Mathematics Series, pages 1–48. AMS/IAS/SIAM, 2018.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36:158–183, 2006a.

Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59:80–90, 2016.

Petros Drineas and Michael W Mahoney. Lectures on randomized numerical linear algebra. *arXiv preprint arXiv:1712.08880*, 2017.

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006b.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, December 2012.

Alan Edelman and N Raj Rao. Random matrix theory. *Acta numerica*, 14:233, 2005.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

Milana Gataric, Tengyao Wang, and Richard J Samworth. Sparse principal component analysis via axis-aligned random projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020.

AB Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and D Rubin. Bayesian data analysis third edition. boca raton. *FL: CRC Press*, 2013.

Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5): 2580–2594, 2011a.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011b.

JA Hartigan. Linear bayesian methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 446–454, 1969.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer series in statistics, 2009.

Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv e-prints*, art. arXiv:1610.02527, Oct 2016a.

Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv e-prints*, art. arXiv:1610.05492, Oct 2016b.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373, International Conference on Learning Representations (ICLR) 2020*, 2019.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.

P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911, 2015.

Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. Also available at: arXiv:1104.5557.

Vladimir A Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

Goran Marjanovic and Alfred O Hero. l0 sparse inverse covariance estimation. *IEEE Transactions on Signal Processing*, 63(12):3218–3231, 2015.

Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1231–1239. 2009.

Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 456–464, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 91–100, New York, NY, USA, 2013. ACM.

Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. *arXiv preprint arXiv:1504.05477*, 2015.

Jelani Nelson and Huy L. Nguyên. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, pages 117–126, Washington, DC, USA, 2013. IEEE Computer Society.

Jelani Nelson and Huy L Nguyen. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*, pages 883–894. Springer, 2014.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

REAC Paley and Antoni Zygmund. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 266–272. Cambridge University Press, 1932.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1): 1842–1879, 2016.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

Sashank J. Reddi, Jakub Konecný, Peter Richtárik, Barnabás Póczós, and Alex Smola. AIDE: Fast and Communication Efficient Distributed Optimization. *arXiv e-prints*, art. arXiv:1608.06879, Aug 2016.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1000–1008, Bejing, China, 22–24 Jun 2014. PMLR.

Radhendushka Srivastava, Ping Li, and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3): 954–970, 2016.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4): 1454–1485, 2017.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18:1–50, 2018a.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. GIANT: Globally improved approximate Newton method for distributed optimization. In *Advances in Neural Information Processing Systems*, pages 2332–2342, 2018b.

Greg Welch and Gary Bishop. An introduction to the kalman filter, 1995.

Virginia Vassilevska Williams. Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 887–898, 2012.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

Fan Yang, Sifan Liu, Edgar Dobriban, and David P Woodruff. How to reduce dimension with pca and random projections? *arXiv preprint arXiv:2005.00511*, 2020.

Z. Yao, A. Gholami, K. Keutzer, and M. W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. Technical Report Preprint: arXiv:1912.07145, 2019.

Z. Yao, A. Gholami, S. Shen, K. Keutzer, and M. W. Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. Technical Report Preprint: arXiv:2006.00719, 2020.

Raphael Yuster and Uri Zwick. Fast sparse matrix multiplication. *ACM Transactions On Algorithms (TALG)*, 1(1):2–13, 2005.

Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 362–370, Lille, France, 07–09 Jul 2015. PMLR.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14(1):3321–3363, January 2013.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. Parallelized stochastic gradient descent. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2595–2603. Curran Associates, Inc., 2010.

Marko Znidaric. Asymptotic expansion for inverse moments of binomial and poisson distributions. *The Open Mathematics, Statistics and Probability Journal*, 1(1), 2009.

Appendix A. Preliminaries

Notations. In the remainder of the article, we follow the convention of denoting scalars by lowercase, vectors by lowercase boldface, and matrices by uppercase boldface letters. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the spectral or operator norm for matrices, and $\|\cdot\|_F$ is the Frobenius norm for matrices. For vector $\mathbf{v} \in \mathbb{R}^d$, we let $\|\mathbf{v}\|_1 := \sum_{i=1}^d |v_i|$ denote the ℓ_1 norm and $\|\mathbf{v}\|_\infty := \max_i |v_i|$ denote the ℓ_∞ norm of $\mathbf{v}$. We use $\lambda_{\max}(\mathbf{A})$ to denote the maximum eigenvalue of a symmetric matrix $\mathbf{A}$. We say $\mathbf{A} \preceq \mathbf{B}$ if and only if $\mathbf{B} - \mathbf{A}$ is positive semi-definite. We use $\mathbf{A} \circ \mathbf{B}$ to denote the entry-wise Hadamard product of matrices or vectors. For random vectors or matrices, we say $\mathbf{A} \overset{d}{=} \mathbf{B}$ if $\mathbf{A}$ follows the same distribution as $\mathbf{B}$. For positive semi-definite (p.s.d.) matrices $\mathbf{A}$ and $\mathbf{B}$, or non-negative scalars a and b, we use $\mathbf{A} \approx_\eta \mathbf{B}$ and $a \approx_\eta b$ to denote the relative error approximation (Definition 1). The big-O notation is used to absorb constant factors in upper bounds, where the constant only depends on other big-O constants appearing in a given statement (thus, all constants can be made absolute).

An important linear algebraic result that will be used in proving the restricted Bai-Silverstein inequality (Theorem 14) is the following Perron-Frobenius theorem on non-negative matrices. While the most well known version of the Perron-Frobenius theorem concerns matrices with strictly positive entries, there is also a version for matrices with only non-negative entries.

Lemma 17 (Perron-Frobenius theorem, (Meyer, 2000, claims 8.3.1 and 8.3.2)) *For a non-negative symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ such that $[\mathbf{A}]_{ij} \geq 0$ for all $i, j \in \{1, \ldots, n\}$, then the largest eigenvalue of $\mathbf{A}$ is non-negative, i.e., $r = \lambda_{\max}(\mathbf{A}) \geq 0$. Moreover, there is a corresponding eigenvector $\mathbf{z}$, i.e., $\mathbf{A}\mathbf{z} = r\mathbf{z}$ with non-negative entries $\mathbf{z}_i \geq 0$ for all i.*

Our analysis of the inversion bias (proof of Theorem 11) crucially relies on a standard rank-one update formula for the matrix inverse, which is given below.

Lemma 18 (Sherman-Morrison formula) *For an invertible matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u} \neq 0$. If this holds, then*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}.$$

In particular, it follows that:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}.$$

Our proofs rely on different types of concentration and anti-concentration inequalities, from scalars to quadratic forms of the type $\mathbf{x}^\top\mathbf{B}\mathbf{x}$, and eventually to matrix concentration bounds. These technical lemmas are collected in this section and will be repeatedly used in the proofs of our main results.

A.1. Scalar concentration and anti-concentration inequalities

The Burkholder inequality (Burkholder, 1973) provides moment bounds on the sum of a martingale difference sequence. It is used to show Lemma 25 as part of the proof of Theorem 11.

Lemma 19 (Burkholder inequality, (Burkholder, 1973)) *For $\{x_j\}_{j=1}^m$ a real martingale difference sequence with respect to the increasing σ field $\mathcal{F}_j$, we have, for $L > 1$, there exists $C_L > 0$ such that*

$$\mathbb{E}\Big[\Big|\sum_{j=1}^m x_j\Big|^L\Big] \le C_L \cdot \mathbb{E}\Big[\Big(\sum_{j=1}^m |x_j|^2\Big)^{L/2}\Big].$$

The Paley-Zygmund inequality is used to establish an anti-concentration inequality for the Binomial distribution (Lemma 36), which is the key in deriving a lower bound for the inversion bias of leverage score sampling in Appendix F.

Lemma 20 (Paley-Zygmund inequality, (Paley and Zygmund, 1932)) *For any non-negative variable Z with finite variance and $\theta \in (0,1)$, we have:*

$$\Pr\big(Z \ge \theta\,\mathbb{E}[Z]\big) \ge (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

A.2. Quadratic form concentration

Being the key object of (one of) the structural conditions in Theorem 11, the (random) quadratic form of the type $\mathbf{x}^\top \mathbf{B}\mathbf{x}$ will consistently appear in our analysis, for instance in the form of the Bai-Silverstein inequality in Lemma 13 on quadratic form variance, as well as the following Hanson-Wright inequality on the tail probability.

Lemma 21 (Hanson-Wright inequality, (Rudelson and Vershynin, 2013, Theorem 1.1)) *Let $\mathbf{x}$ have independent $O(1)$-sub-gaussian entries with mean zero and unit variance. Then, there is $c = \Omega(1)$ such that for any $n \times n$ matrix $\mathbf{B}$ and $t \ge 0$,*

$$\Pr\Big\{|\mathbf{x}^\top \mathbf{B}\mathbf{x} - \mathrm{tr}(\mathbf{B})| \ge t\Big\} \le 2\exp\left(-c\min\left\{\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|}\right\}\right).$$

A.3. Matrix concentration inequalities

When random matrices are considered, different variants of Matrix Chernoff/Bernstein inequalities will be needed to handle the case where the random matrix under study is known to have (almost surely) bounded operator norm, or only to admit a subexponential decay for its higher order moments.

Lemma 22 (Matrix Bernstein: Bounded Case, (Tropp, 2012, Theorem 1.4)) *For $i = 1, 2, ...$, consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{X}_i] = \mathbf{0}, \quad \lambda_{\max}(\mathbf{X}_i) \le R \quad \textit{almost surely.}$$

Then, defining the variance parameter $\sigma^2 = \|\sum_i \mathbb{E}[\mathbf{X}_i^2]\|$, for any $t > 0$ we have:

$$\Pr\Big\{\lambda_{\max}\Big(\sum\nolimits_i \mathbf{X}_i\Big) \ge t\Big\} \le d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Lemma 23 (Matrix Bernstein: Subexponential Case, (Tropp, 2012, Theorem 6.2)) *For $i = 1, 2, ...,$ consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{X}_i] = \mathbf{0}, \quad \mathbb{E}[\mathbf{X}_i^p] \preceq \frac{p!}{2} \cdot R^{p-2}\mathbf{A}_i^2 \quad \textit{for} \quad p = 2, 3, ...$$

Then, defining the variance parameter $\sigma^2 = \| \sum_i \mathbf{A}_i^2 \|$, for any $t > 0$ we have:

$$\Pr\Big\{\lambda_{\max}\Big(\sum\nolimits_i \mathbf{X}_i\Big) \geq t\Big\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt}\right).$$

Lemma 24 (Matrix Chernoff, (Tropp, 2012, Theorem 1.1 and Remark 5.3)) *For $i = 1, 2, ...,$ consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent positive semi-definite random matrices such that $\mathbb{E}\big[\sum_i \mathbf{X}_i\big] = \mathbf{I}$ and $\|\mathbf{X}_i\| \leq R$. Then, for any $t \geq \mathrm{e}$, we have:*

$$\Pr\Big\{\Big\|\sum_i \mathbf{X}_i\Big\| \geq t\Big\} \leq d \cdot \left(\frac{\mathrm{e}}{t}\right)^{t/R}.$$

Appendix B. Structural conditions for small inversion bias

In this section, we prove Theorem 11, which gives two structural conditions for a random sketch of a rank d matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ to have small inversion bias. We assume that the sketching matrix $\mathbf{S}_m \in \mathbb{R}^{m \times n}$ consists of $m \geq 8d$ i.i.d. rows $\frac{1}{\sqrt{m}}\mathbf{x}_i^\top$, where $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{I}$. To simplify the analysis, we assume that m is divisible by 3.

B.1. Proof of Theorem 11

Note that the subspace embedding assumption (based on Condition 1) immediately implies the result with $\epsilon = O(1)$, so without loss of generality we can assume that $\alpha\sqrt{d}/m \leq 1$. Let $\mathbf{H} = \mathbf{A}^\top\mathbf{A}$ and $\mathbf{Q} = (\gamma\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A})^{-1}$ for $\gamma = \frac{m}{m-d}$. Moreover, let $\mathbf{S}_{-i}$ denote $\mathbf{S}_m$ without the ith row, with $\mathbf{Q}_{-i} = (\gamma\mathbf{A}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{A})^{-1}$. Finally, for $t = m/3$, we define the following events:

$$\mathcal{E}_j : \ \frac{1}{t}\mathbf{A}^\top\Big(\sum_{i=t(j-1)+1}^{tj} \mathbf{x}_i\mathbf{x}_i^\top\Big)\mathbf{A} \succeq \frac{1}{2}\cdot\mathbf{A}^\top\mathbf{A}, \quad j = 1, 2, 3, \qquad \mathcal{E} = \bigwedge_{j=1}^{3}\mathcal{E}_j. \tag{2}$$

For each j, the meaning of the event $\mathcal{E}_j$ is that the average of the rank one matrices $\mathbf{x}_i\mathbf{x}_i^\top$ over the corresponding j-th third of indices $1, \ldots, m$ forms a sketch for $\mathbf{A}$ that is a "lower" spectral approximation of $\mathbf{A}^\top\mathbf{A}$.

Note that events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ are independent, and for each $i \in \{1, ..., m\}$ there is a $j = j(i) \in \{1, 2, 3\}$ such that:

1. $\mathcal{E}_j$ is independent of $\mathbf{x}_i$; and
2. $\mathcal{E}_j$ implies that $\mathbf{Q}_{-i} \preceq \gamma\mathbf{Q}_{-i} = (\mathbf{A}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{A})^{-1} \preceq 6\cdot(\mathbf{A}^\top\mathbf{A})^{-1} = 6\cdot\mathbf{H}^{-1}$.

Here we use that $\mathbf{A}^\top\mathbf{S}_m^\top\mathbf{S}_m\mathbf{A}$ is the average of the three matrices to which the conditions in $\mathcal{E}_j$ refer to, and also that $m \geq 2d$.

From the subspace embedding assumption and the union bound we conclude that $\Pr(\mathcal{E}) \geq 1-\delta$. Letting $\mathbb{E}_{\mathcal{E}}$ denote the expectation conditioned on $\mathcal{E}$ and $\gamma_i = 1 + \frac{\gamma}{m}\mathbf{x}_i^\top \mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i$, we have:

$$\begin{aligned}
\mathbf{I} - \mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H} &= -\mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H} + \gamma\,\mathbb{E}_{\mathcal{E}}[\mathbf{Q}\mathbf{A}^\top \mathbf{S}_m^\top \mathbf{S}_m \mathbf{A}] = -\mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H} + \gamma\,\mathbb{E}_{\mathcal{E}}[\mathbf{Q}\mathbf{A}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{A}] \\
&\overset{(*)}{=} -\mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H} + \gamma\,\mathbb{E}_{\mathcal{E}}\Big[\tfrac{\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{A}}{1+\frac{\gamma}{m}\mathbf{x}_i^\top \mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i}\Big] \\
&= -\mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H} + \mathbb{E}_{\mathcal{E}}[\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{A}] + \mathbb{E}_{\mathcal{E}}\big[(\tfrac{\gamma}{\gamma_i}-1)\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{A}\big] \\
&= \underbrace{\mathbb{E}_{\mathcal{E}}[\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}]}_{\mathbf{Z}_0} + \underbrace{\mathbb{E}_{\mathcal{E}}[\mathbf{Q}_{-i} - \mathbf{Q}]\mathbf{H}}_{\mathbf{Z}_1} + \underbrace{\mathbb{E}_{\mathcal{E}}\big[(\tfrac{\gamma}{\gamma_i}-1)\mathbf{Q}_{-i}\mathbf{A}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{A}\big]}_{\mathbf{Z}_2},
\end{aligned}$$

for a fixed i, where $(*)$ uses the Sherman-Morrison rank-one update formula (Lemma 18). We also used the fact that due to symmetry in the definition of event $\mathcal{E}$, the marginal distributions of the random vectors $\mathbf{x}_i$ are identical after conditioning (even though they are no longer independent and identically distributed). To obtain the result, it suffices to bound:

$$\begin{aligned}
\|\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H}^{\frac{1}{2}}\| &= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2)\mathbf{H}^{-\frac{1}{2}}\| \\
&\leq \|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_0\mathbf{H}^{-\frac{1}{2}}\| + \|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_1\mathbf{H}^{-\frac{1}{2}}\| + \|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_2\mathbf{H}^{-\frac{1}{2}}\|. \qquad (3)
\end{aligned}$$

We start by bounding the first term. Without loss of generality, assume that events $\mathcal{E}_1$ and $\mathcal{E}_2$ are both independent of $\mathbf{x}_i$, and let $\mathcal{E}' = \mathcal{E}_1 \wedge \mathcal{E}_2$ as well as $\delta_3 = \Pr(\neg\mathcal{E}_3)$. We have:

$$\begin{aligned}
\mathbf{Z}_0 &= \frac{1}{1-\delta_3}\cdot\Big(\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}] - \mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}\cdot \mathbf{1}_{\neg\mathcal{E}_3}]\Big) \\
&= -\frac{1}{1-\delta_3}\cdot\mathbb{E}_{\mathcal{E}'}\big[\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}\cdot \mathbf{1}_{\neg\mathcal{E}_3}\big].
\end{aligned}$$

Above, we evaluated the expectation $\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}]$ by first conditioning on all randomness except $\mathbf{x}_i$, and using the independence of $\mathbf{x}_i$ and $\mathcal{E}'$, as well as $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}$.

Thus, since $\delta_3 \leq \frac{1}{2}$, we obtain that:

$$\begin{aligned}
\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_0\mathbf{H}^{-\frac{1}{2}}\| &\leq 2\left\|\mathbb{E}_{\mathcal{E}'}\big[\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\cdot \mathbf{1}_{\neg\mathcal{E}_3}\big]\right\| \\
&\leq 2\,\mathbb{E}_{\mathcal{E}'}\Big[\big\|\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\big\|\cdot \mathbf{1}_{\neg\mathcal{E}_3}\Big] \\
&\leq 2\,\mathbb{E}_{\mathcal{E}'}\Big[\|\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}}\|\cdot\big\|\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top(\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I})\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\big\|\cdot \mathbf{1}_{\neg\mathcal{E}_3}\Big] \\
&\leq 12\,\mathbb{E}_{\mathcal{E}'}\Big[\big(\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i + 1\big)\cdot \mathbf{1}_{\neg\mathcal{E}_3}\Big].
\end{aligned}$$

Note that $\mathbb{E}[\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i] = d$, and using Condition 2 (Restricted Bai-Silverstein), we have $\mathrm{Var}[\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i] \leq \alpha\cdot d$ (and both are still true after conditioning on $\mathcal{E}'$, because it is independent of $\mathbf{x}_i$). Chebyshev's inequality thus implies that for $x \geq 2d$ we have $\Pr(\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i \geq x \mid \mathcal{E}') \leq C\alpha d/x^2$. Combining this with the assumption that $\delta_3 \leq \delta \leq 1/m^3$, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}'}\big[\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i\cdot \mathbf{1}_{\neg\mathcal{E}}\big] &= \int_0^\infty \Pr(\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i\cdot \mathbf{1}_{\neg\mathcal{E}} \geq x \mid \mathcal{E}')\,dx \\
&\leq 2m^2\delta_3 + \int_{2m^2}^\infty \Pr(\mathbf{x}_i^\top \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top \mathbf{x}_i \geq x)\,dx \\
&\leq \frac{2}{m} + C\alpha d\int_{2m^2}^\infty \frac{1}{x^2}\,dx \leq \frac{2}{m} + C\frac{\alpha d}{m^2},
\end{aligned}$$

which implies that $\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_0\mathbf{H}^{-\frac{1}{2}}\| = O(1/m+\alpha d/m^2) = O(\alpha\sqrt{d}/m)$. We now move on to bounding the second term in (3). In the following, we will use the observation that for a p.s.d. random matrix $\mathbf{C}$ (or non-negative random variable) in the probability space of $\mathbf{S}_m$, we have:

$$\mathbb{E}_{\mathcal{E}}[\mathbf{C}] = \frac{\mathbb{E}[(\prod_{j=1}^{3}\mathbf{1}_{\mathcal{E}_j})\cdot\mathbf{C}]}{\Pr(\mathcal{E})} \preceq \frac{1}{1-\delta}\,\mathbb{E}[\mathbf{1}_{\mathcal{E}'}\cdot\mathbf{C}] \preceq 2\cdot\mathbb{E}_{\mathcal{E}'}[\mathbf{C}]. \tag{4}$$

Using the above, and the fact that event $\mathcal{E}'$ is independent of $\mathbf{x}_i$, we have:

$$\mathbb{E}_{\mathcal{E}}[\mathbf{Q}_{-i}-\mathbf{Q}] \preceq 2\cdot\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}-\mathbf{Q}] = \frac{2\gamma}{m}\cdot\mathbb{E}_{\mathcal{E}'}\left[\gamma_i^{-1}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\right] \preceq \frac{2\gamma}{m}\cdot\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}\mathbf{H}\mathbf{Q}_{-i}].$$

We now bound the second term in (3) by using the fact that $\mathcal{E}'$ implies $\mathbf{H}^{\frac{1}{2}}\gamma\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}} \preceq 6\mathbf{I}$:

$$\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_1\mathbf{H}^{-\frac{1}{2}}\| = \|\mathbf{H}^{\frac{1}{2}}\mathbb{E}_{\mathcal{E}}[\mathbf{Q}_{-i}-\mathbf{Q}]\mathbf{H}^{\frac{1}{2}}\| \le \frac{2\gamma}{m}\cdot\mathbb{E}_{\mathcal{E}'}\left[\|\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}}\cdot\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}}\|\right] \le \frac{2}{m}\cdot 36.$$

We next bound the last term in (3), applying the Cauchy-Schwarz inequality twice:

$$\begin{aligned}
\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_2\mathbf{H}^{-\frac{1}{2}}\| &= \sup_{\|\mathbf{v}\|=1,\ \|\mathbf{u}\|=1}\mathbb{E}_{\mathcal{E}}\left[\left|\tfrac{\gamma}{\gamma_i}-1\right|\cdot\mathbf{v}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\mathbf{u}\right] \\
&\le \sqrt{\mathbb{E}_{\mathcal{E}}\left[(\tfrac{\gamma}{\gamma_i}-1)^2\right]}\cdot\sup_{\|\mathbf{v}\|=1,\ \|\mathbf{u}\|=1}\sqrt{\mathbb{E}_{\mathcal{E}}\left[(\mathbf{v}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\cdot\mathbf{x}_i^\top\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\mathbf{u})^2\right]} \\
&\le \underbrace{\sqrt{\mathbb{E}_{\mathcal{E}}\left[(\gamma_i-\gamma)^2\right]}}_{T_1}\cdot\underbrace{\sup_{\|\mathbf{u}\|=1}\sqrt[4]{\mathbb{E}_{\mathcal{E}}\left[(\mathbf{u}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i)^4\right]}}_{T_2}\cdot\underbrace{\sup_{\|\mathbf{u}\|=1}\sqrt[4]{\mathbb{E}_{\mathcal{E}}\left[(\mathbf{u}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top\mathbf{x}_i)^4\right]}}_{T_3}.
\end{aligned}$$

To bound T_3, we rely on Restricted Bai-Silverstein with $\mathbf{B} = \mathbf{A}\mathbf{H}^{-\frac{1}{2}}\mathbf{u}\mathbf{u}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top$, noting that $\mathrm{tr}(\mathbf{B}^2) = \mathrm{tr}(\mathbf{B}) = (\mathbf{u}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{H}\mathbf{H}^{-\frac{1}{2}}\mathbf{u})^2 = \|\mathbf{u}\|^4 = 1$. Recall that event $\mathcal{E}'$ is independent of $\mathbf{x}_i$, so we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}}\left[(\mathbf{u}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top\mathbf{x}_i)^4\right] &\le 2\,\mathbb{E}_{\mathcal{E}'}\left[(\mathbf{u}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top\mathbf{x}_i)^4\right] \\
&= 2\,\mathbb{E}\left[(\mathbf{x}_i^\top\mathbf{B}\mathbf{x}_i)^2\right] \\
&= 2\mathrm{Var}[\mathbf{x}_i^\top\mathbf{B}\mathbf{x}_i] + 2\big(\mathbb{E}[\mathbf{x}_i^\top\mathbf{B}\mathbf{x}_i]\big)^2 \\
&\le 2\alpha\cdot\mathrm{tr}(\mathbf{B}^2) + 2\big(\mathrm{tr}(\mathbf{B})\big)^2 = 2(\alpha+1),
\end{aligned}$$

obtaining that $T_3 = O(\sqrt[4]{\alpha+1})$. We can similarly bound T_2 by letting $\mathbf{B} = \mathbf{A}\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}}\mathbf{u}\mathbf{u}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{A}^\top$. Note that, conditioned on $\mathcal{E}'$, we again have

$$\mathrm{tr}(\mathbf{B}^2) = \big(\mathbf{u}^\top(\mathbf{H}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{H}^{\frac{1}{2}})^2\mathbf{u}\big)^2 \le 6^4,$$

so analogously as above we conclude that $T_2 = O(\sqrt[4]{\alpha+1})$.

It thus remains to bound the term T_1. First, note that:

$$\mathbb{E}_{\mathcal{E}}[(\gamma-\gamma_i)^2] \le 2\,\mathbb{E}_{\mathcal{E}'}[(\gamma-\gamma_i)^2] = 2\,(\gamma-\bar{\gamma})^2 + 2\,\mathbb{E}_{\mathcal{E}'}[(\gamma_i-\bar{\gamma})^2], \tag{5}$$

where $\bar{\gamma} = \mathbb{E}_{\mathcal{E}'}[\gamma_i] = 1 + \frac{\gamma}{m}\mathrm{tr}(\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}]\mathbf{H})$. To bound the second term in (5), we write:

$$\mathbb{E}_{\mathcal{E}'}[(\gamma_i - \bar{\gamma})^2] = \frac{\gamma^2}{m^2}\,\mathbb{E}_{\mathcal{E}'}\Big[\big(\mathrm{tr}(\mathbf{Q}_{-i} - \mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}])\mathbf{H}\big)^2\Big] + \frac{\gamma^2}{m^2}\,\mathbb{E}_{\mathcal{E}'}\Big[\big(\mathrm{tr}(\mathbf{Q}_{-i}\mathbf{H}) - \mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big)^2\Big].$$

The latter term can be bounded by again using Condition 2, with $\mathbf{B} = \mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top$, obtaining:

$$\frac{\gamma^2}{m^2}\,\mathbb{E}_{\mathcal{E}'}\Big[\big(\mathrm{tr}(\mathbf{Q}_{-i}\mathbf{H}) - \mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big)^2\Big] \le \frac{1}{m^2}\,\mathbb{E}_{\mathcal{E}'}\big[\alpha\cdot\mathrm{tr}((\gamma\mathbf{Q}_{-i}\mathbf{H})^2)\big] \le 36\cdot\frac{\alpha d}{m^2}.$$

The former term can be bounded using the Burkholder inequality for martingale difference sequences. We state this bound as a lemma, proven separately in Appendix B.2.

Lemma 25 *Let $\mathrm{Var}_{\mathcal{E}'}[\cdot]$ be the conditional variance with respect to event $\mathcal{E}' = \mathcal{E}_1 \wedge \mathcal{E}_2$, see (2), with $\mathbf{x}_i$ independent of $\mathcal{E}'$. Then, there is an absolute constant $C > 0$ such that:*

$$\mathrm{Var}_{\mathcal{E}'}\big[\mathrm{tr}(\mathbf{Q}_{-i}\mathbf{H})\big] \le C\cdot d.$$

Using Lemma 25, we conclude that $\mathbb{E}_{\mathcal{E}'}[(\gamma_i - \bar{\gamma})^2] \le C'\cdot\alpha d/m^2$ for some absolute constant C'. It remains to bound the term:

$$|\gamma - \bar{\gamma}| = \left|\frac{m}{m-d} - \Big(1 + \frac{\gamma}{m}\mathrm{tr}(\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}]\mathbf{H})\Big)\right| = \frac{|d - \mathrm{tr}(\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}]\mathbf{H})|}{m-d}.$$

Observe that we have:

$$\begin{aligned}\big|d - \mathrm{tr}\,\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}]\mathbf{H}\big| &= \big|\mathrm{tr}((\mathbb{E}_{\mathcal{E}}[\mathbf{Q}] - \mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}])\mathbf{H}) + \mathrm{tr}(\mathbf{I} - \mathbb{E}_{\mathcal{E}}[\mathbf{Q}]\mathbf{H})\,\big|\\ &= \big|\mathrm{tr}((\mathbb{E}_{\mathcal{E}} - \mathbb{E}_{\mathcal{E}'})[\mathbf{Q}_{-i}]\mathbf{H}) + \mathrm{tr}(-\mathbf{Z}_1) + \mathrm{tr}(\mathbf{Z}_0 + \mathbf{Z}_1 + \mathbf{Z}_2)\big|\\ &\le \big|\mathrm{tr}((\mathbb{E}_{\mathcal{E}} - \mathbb{E}_{\mathcal{E}'})[\mathbf{Q}_{-i}]\mathbf{H})\big| + |\mathrm{tr}(\mathbf{Z}_0)| + |\mathrm{tr}(\mathbf{Z}_2)|.\end{aligned}$$

The first two terms can be bounded similarly as we did $\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_0\mathbf{H}^{-\frac{1}{2}}\|$, obtaining that $|\mathrm{tr}(\mathbf{Z}_0)| = O(\alpha d/m)$, and also:

$$\big|\mathrm{tr}((\mathbb{E}_{\mathcal{E}} - \mathbb{E}_{\mathcal{E}'})[\mathbf{Q}_{-i}]\mathbf{H})\big| = \frac{\delta_3}{1-\delta_3}\big|\mathrm{tr}((\mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}] - \mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i} \mid \neg\mathcal{E}_3])\mathbf{H})\big| = O(d\delta_3) = O(d/m^3).$$

For the last term, we have:

$$\begin{aligned}|\mathrm{tr}(\mathbf{Z}_2)| &= \Big|\mathbb{E}_{\mathcal{E}}\big[(\tfrac{\gamma}{\gamma_i} - 1)\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big]\Big|\\ &\le \Big|\mathbb{E}_{\mathcal{E}}\big[\tfrac{\gamma-\bar{\gamma}}{\gamma_i}\,\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big]\Big| + \Big|\mathbb{E}_{\mathcal{E}}\big[\tfrac{\bar{\gamma}-\gamma_i}{\gamma_i}\,\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big]\Big|\\ &\le |\gamma - \bar{\gamma}|\cdot\mathbb{E}_{\mathcal{E}}[\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i] + (m-d)\cdot\mathbb{E}\big[|\gamma_i - \bar{\gamma}|\big]\\ &\le |\gamma - \bar{\gamma}|\cdot\frac{6}{1-\delta}\,\mathbb{E}_{\mathcal{E}'}[\mathbf{x}_i^\top\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top\mathbf{x}_i] + (m-d)\cdot\sqrt{\mathbb{E}[(\gamma_i - \bar{\gamma})^2]}\\ &\le |\gamma - \bar{\gamma}|\cdot\frac{6}{1-\delta}\,d + \sqrt{C'\alpha d}.\end{aligned}$$

The bound for the second term $\Big|\mathbb{E}_{\mathcal{E}}\big[\frac{\bar{\gamma}-\gamma_i}{\gamma_i}\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i\big]\Big|$ comes from the definition of $\gamma_i = 1 + \frac{\gamma}{m}\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i$, because $\mathbf{x}_i^\top\mathbf{A}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{x}_i/\gamma_i \le m/\gamma = m - d$.

Thus, putting this together we conclude that:

$$|\gamma-\bar{\gamma}|\cdot\big(1-\tfrac{6d}{(m-d)(1-\delta)}\big)\leq O(\alpha d/m^2)+O(\sqrt{\alpha d}/m)=O(\sqrt{\alpha d}/m),$$

which for $m\geq 8d$ and $\delta\leq 1/m^3$ implies that $(\gamma-\bar{\gamma})^2=O(\alpha d/m^2)$ so we get $T_1=O(\sqrt{\alpha d}/m)$. Finally, we obtain the bound $\|\mathbf{H}^{\frac{1}{2}}\mathbf{Z}_2\mathbf{H}^{-\frac{1}{2}}\|\leq T_1\cdot T_2\cdot T_3=O(\alpha\sqrt{d}/m)$, which concludes the proof.

B.2. Proof of Lemma 25

Let $\mathbf{Q}_{-ij}$ denote the matrix $(\gamma\mathbf{A}^\top\mathbf{S}_{-ij}^\top\mathbf{S}_{-ij}\mathbf{A})^{-1}$ where $\mathbf{S}_{-ij}$ is the matrix $\mathbf{S}_m$ without the ith and jth rows and $\gamma=\frac{m}{m-d}$. Let $\mathbb{E}_{\mathcal{E}',j}[\cdot]$ be the conditional expectation with respect to $\mathcal{E}'$ and the σ-field $\mathcal{F}_j$ generating the rows $\frac{1}{\sqrt{m}}\mathbf{x}_1^\top\ldots,\frac{1}{\sqrt{m}}\mathbf{x}_j^\top$ of $\mathbf{S}$. First note that

$$\begin{aligned}\mathrm{tr}(\mathbf{Q}_{-i}-\mathbb{E}_{\mathcal{E}'}\mathbf{Q}_{-i})\mathbf{A}^\top\mathbf{A}&=\mathbb{E}_{\mathcal{E}',m}[\mathrm{tr}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{A}]-\mathbb{E}_{\mathcal{E}',0}[\mathrm{tr}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{A}]\\&=\sum_{j=1}^m\big(\mathbb{E}_{\mathcal{E}',j}[\mathrm{tr}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{A}]-\mathbb{E}_{\mathcal{E}',j-1}[\mathrm{tr}\mathbf{Q}_{-i}\mathbf{A}^\top\mathbf{A}]\big)=-\sum_{j=1}^m(\psi_j+\xi_j),\\ \text{where}\quad\psi_j&=(\mathbb{E}_{\mathcal{E}',j}-\mathbb{E}_{\mathcal{E}',j-1})[\mathrm{tr}(\mathbf{Q}_{-ij}-\mathbf{Q}_{-i})\mathbf{A}^\top\mathbf{A}]\\ \text{and}\ \ \xi_j&=-(\mathbb{E}_{\mathcal{E}',j}-\mathbb{E}_{\mathcal{E}',j-1})[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}].\end{aligned}$$

This forms a martingale difference sequence and hence falls within the scope of the Burkholder inequality (Burkholder, 1973), recalled as follows. We mention that similar martingale decomposition techniques are common in random matrix theory, see e.g., (Bai and Silverstein, 2010). Also, for the case $L=2$ that we will use, Burkholder inequality is nothing but the law of iterated variance.

Lemma 26 (Burkholder (1973)) *For $\{x_j\}_{j=1}^m$ a real martingale difference sequence with respect to the increasing σ field $\mathcal{F}_j$, we have, for $L>1$, there exists $C_L>0$ such that*

$$\mathbb{E}\bigg[\Big|\sum_{j=1}^m x_j\Big|^L\bigg]\leq C_L\cdot\mathbb{E}\bigg[\Big(\sum_{j=1}^m|x_j|^2\Big)^{L/2}\bigg].$$

Note that for each pair i,j, one of $\mathcal{E}_1,\mathcal{E}_2$ is independent of both $\mathbf{x}_i$ and $\mathbf{x}_j$. Without loss of generality, suppose that this is $\mathcal{E}_1$. Then, in particular, $\mathcal{E}_1$ implies that $\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\preceq 6\,\mathbf{I}$. Thus, conditioned on $\mathcal{E}_1$, it follows that

$$\begin{aligned}\mathrm{tr}(\mathbf{Q}_{-ij}-\mathbf{Q}_{-i})\mathbf{A}^\top\mathbf{A}&=\mathrm{tr}\bigg(\frac{\frac{\gamma}{m}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}}{1+\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}\mathbf{A}^\top\mathbf{A}\bigg)\\&=\frac{\frac{\gamma}{m}\mathbf{x}_j^\top(\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top)^2\mathbf{x}_j}{1+\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}\leq\frac{6\cdot\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}{1+\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}\leq 6,\end{aligned}$$

which implies that $|\psi_j|\leq 6$. We now provide a bound on the second moment of ψ_j, bounding the $\mathcal{E}'$-conditional expectation in terms of the $\mathcal{E}_1$-conditional expectation analogously as in (4):

$$\begin{aligned}\mathbb{E}_{\mathcal{E}'}[\psi_j^2]&\leq 2\cdot\mathbb{E}_{\mathcal{E}_1}\Bigg[\bigg(\frac{6\cdot\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}{1+\frac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j}\bigg)^2\Bigg]\leq 72\cdot\mathbb{E}_{\mathcal{E}_1}[\tfrac{\gamma}{m}\mathbf{x}_j^\top\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{x}_j]\\&=72\cdot\frac{\mathbb{E}_{\mathcal{E}_1}[\mathrm{tr}\,\mathbf{A}\mathbf{Q}_{-ij}\mathbf{A}^\top]}{m-d}\leq 72\cdot 6\cdot\frac{d}{m-d}.\end{aligned}$$

We now aim to bound $|\xi_j|$. Since $\mathcal{E}_1$ is independent of $\mathbf{x}_j$, we have $\mathbb{E}_{\mathcal{E}_1,j}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}] = \mathbb{E}_{\mathcal{E}_1,j-1}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}]$. Furthermore, letting $\delta_2 = \Pr(\neg\mathcal{E}_2)$, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{E}_1,j-1}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}] &= \mathbb{E}_{\mathcal{E}',j-1}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}](1-\delta_2) + \mathbb{E}_{\mathcal{E}_1,j-1}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A} \mid \neg\mathcal{E}_2]\delta_2,\\ \mathbb{E}_{\mathcal{E}_1,j}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}] &= \mathbb{E}_{\mathcal{E}',j}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}](1-\delta_2) + \mathbb{E}_{\mathcal{E}_1,j}[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A} \mid \neg\mathcal{E}_2]\delta_2.\end{aligned}$$

Thus, subtracting the two equalities from each other, we conclude that:

$$\begin{aligned}|\xi_j| &= |(\mathbb{E}_{\mathcal{E}',j} - \mathbb{E}_{\mathcal{E}',j-1})[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A}]|\\ &\leq \delta_2 \cdot \frac{|(\mathbb{E}_{\mathcal{E}_1,j} - \mathbb{E}_{\mathcal{E}_1,j-1})[\mathrm{tr}\,\mathbf{Q}_{-ij}\mathbf{A}^\top\mathbf{A} \mid \neg\mathcal{E}_2]|}{1-\delta_2}\\ &\leq 2\delta_2 \cdot 6d \leq 12 \cdot d/m, \quad \text{for} \quad \delta_2 \leq 1/m.\end{aligned}$$

So, with $x_j = \psi_j + \xi_j$ and $X = -\mathrm{tr}(\mathbf{Q}_{-i} - \mathbb{E}_{\mathcal{E}'}[\mathbf{Q}_{-i}])\mathbf{A}^\top\mathbf{A}$ in Lemma 26, for $L = 2$ we get:

$$\begin{aligned}\mathbb{E}_{\mathcal{E}'}[X^2] &\leq C_2 \cdot \sum_j \mathbb{E}_{\mathcal{E}'}\left[(\psi_j + \xi_j)^2\right]\\ &= C_2 \cdot \sum_j \Big(\mathbb{E}_{\mathcal{E}'}[\psi_j^2] + 2\,\mathbb{E}_{\mathcal{E}'}[\psi_j\xi_j] + \mathbb{E}_{\mathcal{E}'}[\xi_j^2]\Big)\\ &\leq C_2 m \cdot \Big(72 \cdot 6 \cdot \frac{d}{m-d} + 2 \cdot 6 \cdot 12 \cdot \frac{d}{m} + 12^2\frac{d^2}{m^2}\Big) \leq Cd,\end{aligned}$$

where we also used that $m \geq 8d$, thus concluding the proof.

Appendix C. Restricted Bai-Silverstein inequality

In this section, we prove Theorem 14. Specifically, we study Condition 2 (Restricted Bai-Silverstein), the second structural condition for small inversion bias in Theorem 11, which describes the deviation of a quadratic form $\mathbf{x}^\top\mathbf{B}\mathbf{x}$ from its mean, for a random vector $\mathbf{x}$. We start by showing Theorem 14, a generalized version of the lemma of Bai and Silverstein (Lemma 13), which applies when $\mathbf{x}$ has independent entries. Then, in Appendix C.2 we show a similar result for a leverage score sparsified vector, constructed as in Definition 6, which has non-independent entries. Finally, in Appendix C.3 we consider random vectors used in other fast sketching methods, and give lower bounds demonstrating why these methods do not provide satisfactory guarantees for Condition 2.

C.1. Proof of Theorem 14

Since the assumptions on $\mathbf{x}$ only depend on the leverage scores of $\mathbf{A}$, and the conclusion is about the variance of a quadratic form, which only depends on the first four moments of the entries of $\mathbf{x}$, we can assume without loss of generality that the distribution of $\mathbf{x}$ only depends on the leverage scores of $\mathbf{A}$. We will prove the claim for such random vectors $\mathbf{x}$.

We start by proving the following result:

Proposition 27 *Let $\mathbf{A}$ be a fixed $n \times d$ matrix with $n \geq d$, and $\mathbf{x}$ be a random vector with independent entries with mean zero and unit variance, whose distribution only depends on the leverage scores of $\mathbf{A}$. Then, Condition 2 (Restricted Bai-Silverstein) for the matrix $\mathbf{A}$ is equivalent to*

$$\lambda_{\max}\left((\mathbf{U} \circ \mathbf{U})^\top\mathbf{D}(\mathbf{U} \circ \mathbf{U})\right) \leq \alpha - 2,$$

where $\mathbf{U}$ *is the* $n \times d$ *matrix of left singular vectors of* $\mathbf{A}$ *and* $\mathbf{D}$ *is the* $n \times n$ *matrix* $\mathbf{D} = \mathrm{diag}(d_k)$*, with* $d_k = \mathbb{E}x_k^4 - 3$.

Proof The Restricted Bai-Silverstein condition is equivalent to having, for all matrices $\mathbf{B}$ of the form $\mathbf{B} = \mathbf{U}\mathbf{M}\mathbf{U}^\top$, where $\mathbf{U}$ is the matrix of left singular vectors of $\mathbf{A}$,

$$\mathrm{Var}[\mathbf{x}^\top \mathbf{B}\mathbf{x}] \le \alpha \cdot \mathrm{tr}(\mathbf{B}^2).$$

Let $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$. Then this is equivalent to

$$\mathrm{Var}[\mathbf{z}^\top \mathbf{M}\mathbf{z}] \le \alpha \cdot \mathrm{tr}(\mathbf{M}^2). \tag{6}$$

First we claim that it is enough to consider diagonal matrices $\mathbf{M}$. Suppose that we have a condition $C(\mathrm{diag}(\mathbf{U}\mathbf{U}^\top), \alpha)$ that guarantees that (6) holds for diagonal matrices $\mathbf{M}_d$, and that depends only on the leverage scores and α. Now, consider a general matrix $\mathbf{M}$, and suppose it has the eigendecomposition $\mathbf{M} = \mathbf{O}\mathbf{M}_d\mathbf{O}^\top$ for a diagonal matrix $\mathbf{M}_d$. We can write the equivalences

$$\begin{aligned}\mathrm{Var}[\mathbf{z}^\top \mathbf{M}\mathbf{z}] &\le \alpha \cdot \mathrm{tr}(\mathbf{M}^2)\\ \mathrm{Var}[\mathbf{z}^\top \mathbf{O}\mathbf{M}_d\mathbf{O}^\top \mathbf{z}] &\le \alpha \cdot \mathrm{tr}([\mathbf{O}\mathbf{M}_d\mathbf{O}^\top]^2)\\ \mathrm{Var}[\mathbf{z}_d^\top \mathbf{M}_d\mathbf{z}_d] &\le \alpha \cdot \mathrm{tr}(\mathbf{M}_d^2)\end{aligned}$$

where $\mathbf{z}_d = \mathbf{O}^\top \mathbf{z} = (\mathbf{U}\mathbf{O})^\top \mathbf{x}$. Now we apply the condition $C(\mathrm{diag}(\mathbf{U}_o\mathbf{U}_o^\top), \alpha)$ to $\mathbf{U}_o = \mathbf{U}\mathbf{O}$ and the diagonal matrix $\mathbf{M}_d$. This condition is applicable, because $\mathbf{M}_o$ is a diagonal matrix, and guarantees (6) for $\mathbf{M}_d$. However, we also have that the row norms of $\mathbf{U}_o = \mathbf{U}\mathbf{O}$ are the same as the row norms of $\mathbf{U}$, because $\mathbf{O}$ simply acts by an orthogonal rotation of the rows. So $\mathrm{diag}(\mathbf{U}_o\mathbf{U}_o^\top) = \mathrm{diag}(\mathbf{U}\mathbf{U}^\top)$. Thus, since the distributions of the sketches we consider only depend on the leverage scores of $\mathbf{A}$, which are the diagonals of the matrix $\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{U}\mathbf{U}^\top$, the condition $C(\mathrm{diag}(\mathbf{U}\mathbf{U}^\top), \alpha)$ guarantees that (6) holds for the original matrix $\mathbf{M}$. This shows that it is enough to establish the condition for diagonal matrices $\mathbf{M}$.

Hence we can rotate $\mathbf{U}$ by the eigenvectors $\mathbf{O}$ of $\mathbf{M}$ into $\mathbf{U}' = \mathbf{U}\mathbf{O}$, and thus assume without loss of generality that $\mathbf{M}$ is diagonal, $\mathbf{M} = \mathrm{diag}(\mathbf{g})$, where $\mathbf{g}$ is a vector. Then, the condition simplifies to

$$\begin{aligned}\mathrm{Var}[\mathbf{z}^\top \mathbf{M}\mathbf{z}] &= \mathrm{Var}[\sum_{i=1}^d z_i^2 g_i]\\ &= \mathbf{g}^\top \mathbf{\Gamma}\mathbf{g} \le \alpha \cdot \|\mathbf{g}\|^2,\end{aligned}$$

where $\mathbf{\Gamma}$ is the covariance matrix of $\mathbf{z} \circ \mathbf{z}$. Here the symbol $\circ$ means entrywise product. This condition has to be true for any vector $\mathbf{g}$. Thus, this condition says exactly that the largest eigenvalue of $\mathbf{\Gamma}$ is at most α:

$$\lambda_{\max}(\mathbf{\Gamma}) \le \alpha.$$

Also we assume that $\mathbb{E}\mathbf{x}\mathbf{x}^\top = \mathbf{I}_m$, hence for any symmetric matrix $\mathbf{F}$ (see e.g., (Bai and Silverstein, 2010; Couillet and Debbah, 2011) and (Mei and Montanari, 2019, Lemma B.6.)),

$$\mathrm{Var}[\mathbf{x}^\top \mathbf{F}\mathbf{x}] = \sum_k d_k F_{kk}^2 + 2\mathrm{tr}(\mathbf{F}^2) \tag{7}$$

where $d_k = \mathbb{E}x_k^4 - 3$. Therefore, applying this for $\mathbf{F} = \mathbf{U}\,\mathrm{diag}(\mathbf{g})\mathbf{U}^\top$, and matching terms, one has $\mathbf{\Gamma} = (\mathbf{U} \circ \mathbf{U})^\top \mathbf{D}\mathbf{U} \circ \mathbf{U} + 2\mathbf{I}_n$, where $\mathbf{D} = \mathrm{diag}(d_k)$ and with $d_k = \mathbb{E}x_k^4 - 3$. This finishes the proof. ■

We now continue with the proof of the main claim (Theorem 14). Based on the above results, as long as the random vector $\mathbf{x}$ has independent entries of zero mean and unit variance, proving Condition 2 boils down to the control of the fourth moment of the distribution.

Let $\mathbf{R} = \mathbf{U} \circ \mathbf{U}$, and let $\mathbf{r}_i$ denote its rows. Note that $\mathbf{r}_i$ have non-negative entries. Let $\mathbf{L} = \mathrm{diag}(1/\|\mathbf{u}_i\|^2) = \mathrm{diag}(1/\|\mathbf{r}_i\|_1) = \mathrm{diag}(1/l_i)$ be the matrix of inverse leverage scores of $\mathbf{A}$, which are also the inverse ℓ_1 norms of the rows $\mathbf{r}_i$ of $\mathbf{R}$. We can simply discard the zero rows to ensure that this is well defined and $\|\mathbf{r}_i\|_1 > 0$ for all indices.

Then if we can bound $\lambda_{\max}(\mathbf{R}^\top \mathbf{L}\mathbf{R}) \le \kappa$, it follows that $\lambda_{\max}(\mathbf{R}^\top \mathbf{D}\mathbf{R}) \le C\kappa \le \alpha - 2$, which is our desired condition as long as α is sufficiently large. We will show this bound with $\kappa = 1$.

Note that $\mathbf{Q} = \mathbf{R}^\top \mathbf{L}\mathbf{R}$ is a symmetric matrix and has non-negative entries, because the rows of $\mathbf{R}$, $\mathbf{r}_i = \mathbf{u}_i \circ \mathbf{u}_i$ are the entry-wise squares of certain vectors, and the entries of $\mathbf{L}$ are all positive. Moreover, it is readily verified that the all ones vector $\mathbf{1}_d$ (which clearly has non-negative entries), is an eigenvalue of $\mathbf{Q}$ with unit eigenvalue,

$$\mathbf{Q}\mathbf{1}_d = \mathbf{1}_d.$$

In other words, $\mathbf{Q}$ is a symmetric doubly stochastic matrix. In more detail, we have

$$\mathbf{Q}\mathbf{1}_d = \mathbf{R}^\top \mathbf{L}\mathbf{R}\mathbf{1}_d = \sum_{i=1}^n \frac{\mathbf{r}_i \mathbf{r}_i^\top}{\|\mathbf{r}_i\|_1}\mathbf{1}_d = \sum_{i=1}^n \mathbf{r}_i \cdot \frac{\mathbf{r}_i^\top \mathbf{1}_d}{\|\mathbf{r}_i\|_1}.$$

Now, clearly, since $\mathbf{r}_i$ have non-negative entries, we have $\mathbf{r}_i^\top \mathbf{1}_d = \|\mathbf{r}_i\|_1$. Therefore, we find

$$\mathbf{Q}\mathbf{1}_d = \sum_{i=1}^n \mathbf{r}_i \cdot \frac{\|\mathbf{r}_i\|_1}{\|\mathbf{r}_i\|_1} = \sum_{i=1}^n \mathbf{r}_i = \mathbf{1}_d.$$

In the last equality, we have used that, since the columns of $\mathbf{U}$ are orthogonal vectors, we have that $\sum_{i=1}^n r_{ij} = 1$ for all $j = 1, \ldots, d$.

Hence, the largest eigenvalue of $\mathbf{Q}$ is at least 1. By the Perron-Frobenius theorem for non-negative matrices, it follows that the largest eigenvalue of $\mathbf{Q}$ is paired with an eigenvector $\mathbf{v}$ of non-negative entries, see e.g., (Meyer, 2000, claims 8.3.1 and 8.3.2). We can write, for any such vector $\mathbf{v} \ge 0$, that

$$\mathbf{Q}\mathbf{v} = \mathbf{R}^\top \mathbf{L}\mathbf{R}\mathbf{v} = \sum_{i=1}^n \frac{\mathbf{r}_i \mathbf{r}_i^\top}{\|\mathbf{r}_i\|_1}\mathbf{v} = \sum_{i=1}^n \mathbf{r}_i \cdot \frac{\mathbf{r}_i^\top \mathbf{v}}{\|\mathbf{r}_i\|_1}.$$

Now, clearly $\mathbf{r}_i^\top \mathbf{v}/\|\mathbf{r}_i\|_1 \le \|\mathbf{v}\|_\infty$. Since each entry of each $\mathbf{r}_i$ is non-negative, we have that $0 \le (\mathbf{Q}\mathbf{v})_j \le (\sum_{i=1}^n r_{ij})\|\mathbf{v}\|_\infty$. As mentioned, we also have that $\sum_{i=1}^n r_{ij} = 1$. Hence,

$$0 \le (\mathbf{Q}\mathbf{v})_j \le \|\mathbf{v}\|_\infty, \; j = 1, \ldots, n.$$

Suppose $\mathbf{v}$ is an eigenvector of $\mathbf{Q}$ with eigenvalue $\lambda \ge 0$, i.e., $\mathbf{Q}\mathbf{v} = \lambda\mathbf{v}$. Based on the above inequality, we find $\|\lambda\mathbf{v}\|_\infty \le \|\mathbf{v}\|_\infty$, hence $\lambda \le 1$. This shows that the largest eigenvalue of $\mathbf{Q}$ is at most unity. Thus, by the above reasoning $\lambda_{\max}(\mathbf{R}^\top \mathbf{D}\mathbf{R}) \le C$, and thus Condition 2 holds as long as $C + 2 \le \alpha$. This finishes the proof.

C.2. Restricted Bai-Silverstein for LESS embeddings

In this section we show that a sub-gaussian random vector sparsified using our leverage score sparsifier (LESS) satisfies Condition 2 (Restricted Bai-Silverstein) with $\alpha = O(1)$. We use this fact later in Appendix D to prove Theorem 8.

Lemma 28 (Restricted Bai-Silverstein for LESS) *Fix a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with rank d and let $\boldsymbol{\xi}$ be a leverage score sparsifier for $\mathbf{A}$. For any p.s.d. matrix $\mathbf{B}$ restricted to the span of $\mathbf{A}$ and any $\mathbf{x}^\top = (x_1, ..., x_n)$ having independent entries with mean zero, unit variance and $\mathbb{E}[\mathbf{x}_i^4] = O(1)$,*

$$\mathrm{Var}\big[(\mathbf{x}\circ\boldsymbol{\xi})^\top\mathbf{B}(\mathbf{x}\circ\boldsymbol{\xi})\big] \leq O(1)\cdot\mathrm{tr}(\mathbf{B}^2).$$

Proof Let $\mathbf{U} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1/2}$ be the orthonormal basis matrix for the span of $\mathbf{A}$, and let $\mathbf{U}_{\boldsymbol{\xi}} = \mathrm{diag}(\boldsymbol{\xi})\mathbf{U}$. Note that $\mathbf{B} = \mathbf{U}\mathbf{U}^\top\mathbf{B}\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{C}\mathbf{U}^\top$ for $\mathbf{C} = \mathbf{U}^\top\mathbf{B}\mathbf{U}$. It follows that:

$$\mathrm{Var}\big[(\mathbf{x}\circ\boldsymbol{\xi})^\top\mathbf{B}(\mathbf{x}\circ\boldsymbol{\xi})\big] = \mathrm{Var}[\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}] = \mathrm{Var}\big[\mathrm{tr}(\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top)\big] + \mathbb{E}\big[\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}]\big],$$

where $\mathrm{Var}_{\boldsymbol{\xi}}$ denotes the conditional variance with respect to $\boldsymbol{\xi}$. Recall that $\xi_i = \sqrt{\frac{b_i}{dp_i}}$, where $b_i = \sum_{t=1}^d 1_{[s_t=i]}$, with s_t sampled i.i.d. from $(p_1, ..., p_n)$ and $p_i \approx_{O(1)} \|\mathbf{u}_i\|^2/d$ (here, $\mathbf{u}_i^\top$ denotes the ith row of $\mathbf{U}$). Thus, $\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{U}_{\boldsymbol{\xi}} = \sum_{t=1}^d \frac{\mathbf{u}_{s_t}\mathbf{u}_{s_t}^\top}{dp_{s_t}}$ and it follows that:

$$\begin{aligned}\mathrm{Var}\big[\mathrm{tr}(\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top)\big] &= \mathrm{Var}\bigg[\sum_{t=1}^d \frac{\mathbf{u}_{s_t}^\top\mathbf{C}\mathbf{u}_{s_t}}{dp_{s_t}}\bigg] = d\,\mathrm{Var}\bigg[\frac{\mathbf{u}_{s_1}^\top\mathbf{C}\mathbf{u}_{s_1}}{dp_{s_1}}\bigg]\\ &\leq d\,\mathbb{E}\bigg[\frac{\mathrm{tr}(\mathbf{C}\mathbf{u}_{s_1}\mathbf{u}_{s_1}^\top\mathbf{C}\mathbf{u}_{s_1}\mathbf{u}_{s_1}^\top)}{d^2p_{s_1}^2}\bigg] \leq \mathbb{E}\bigg[\frac{\|\mathbf{u}_{s_1}\|^2}{dp_{s_1}}\,\frac{\mathbf{u}_{s_1}^\top\mathbf{C}^2\mathbf{u}_{s_1}}{p_{s_1}}\bigg]\\ &\leq O(1)\,\mathbb{E}\bigg[\frac{\mathbf{u}_{s_1}^\top\mathbf{C}^2\mathbf{u}_{s_1}}{p_{s_1}}\bigg] = O(1)\,\mathrm{tr}(\mathbf{U}\mathbf{C}^2\mathbf{U}^\top) = O(1)\,\mathrm{tr}(\mathbf{B}^2).\end{aligned}$$

The Bai-Silverstein inequality (Lemma 13) implies that $\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}] \leq O(1)\cdot\mathrm{tr}\big((\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top)^2\big)$, so we have:

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}]\big] &\leq O(1)\cdot\mathbb{E}\Big[\mathrm{tr}\big((\mathbf{U}_{\boldsymbol{\xi}}\mathbf{C}\mathbf{U}_{\boldsymbol{\xi}}^\top)^2\big)\Big] = O(1)\cdot\mathbb{E}\bigg[\mathrm{tr}\bigg(\Big(\sum_{t=1}^d \frac{\mathbf{C}\mathbf{u}_{s_t}\mathbf{u}_{s_t}^\top}{dp_{s_t}}\Big)^2\bigg)\bigg]\\ &\leq O(1)\sum_{t=1}^d \mathbb{E}\bigg[\frac{\mathrm{tr}(\mathbf{C}\mathbf{u}_{s_t}\mathbf{u}_{s_t}^\top\mathbf{C}\mathbf{u}_{s_t}\mathbf{u}_{s_t}^\top)}{d^2p_{s_t}^2}\bigg] + O(1)\sum_{t\neq r}\mathbb{E}\bigg[\frac{\mathrm{tr}(\mathbf{C}\mathbf{u}_{s_t}\mathbf{u}_{s_t}^\top\mathbf{C}\mathbf{u}_{s_r}\mathbf{u}_{s_r}^\top)}{dp_{s_t}\cdot dp_{s_r}}\bigg]\\ &\leq O(1)\,\mathrm{tr}(\mathbf{B}^2) + O(1)\,\mathrm{tr}\Big(\mathbf{C}\,\mathbb{E}\Big[\frac{\mathbf{u}_{s_1}\mathbf{u}_{s_1}^\top}{p_{s_1}}\Big]\mathbf{C}\,\mathbb{E}\Big[\frac{\mathbf{u}_{s_2}\mathbf{u}_{s_2}^\top}{p_{s_2}}\Big]\Big) \leq O(1)\cdot\mathrm{tr}(\mathbf{B}^2).\end{aligned}$$

Thus, we obtain the desired bound:

$$\mathrm{Var}\big[(\mathbf{x}\circ\boldsymbol{\xi})^\top\mathbf{B}(\mathbf{x}\circ\boldsymbol{\xi})\big] \leq O(1)\cdot\mathrm{tr}(\mathbf{B}^2),$$

which completes the proof. ■

C.3. Lower bounds for other sketching methods

In this section, we show lower bounds for Condition 2 (Restricted Bai-Silverstein) in the context of existing fast sketching techniques. To do that, we first discuss the basic requirement of the framework defined by Theorem 11, namely that the sketching matrix $\mathbf{S}$ must have i.i.d. rows.

Fast sketches with i.i.d. rows. In our discussion, we will focus on three types of sketches: approximate leverage score sampling (Drineas et al., 2006b), Subsampled Randomized Hadamard Transform (Ailon and Chazelle, 2009), and sparse embedding matrices (extensions of the CountSketch (Clarkson and Woodruff, 2017), see (Nelson and Nguyên, 2013; Cohen, 2016)), all of which can be implemented in time nearly linear in the input size. The i.i.d. row assumption can be easily satisfied by any row sampling sketch, including approximate leverage score sampling. The SRHT technically does not satisfy this assumption, however if we treat the Randomized Hadamard Transform as a preprocessing step (given that it does not distort the covariance matrix), then the subsampling part can be analyzed analogously as leverage score sampling. In the case of sparse embedding matrices, the most commonly studied variant has a fixed number of non-zeros per column of $\mathbf{S}$ and so it does not have independent rows, however, it is known that a variant with independently sparsified entries (which fits into the setup of Theorem 11) achieves nearly matching approximation guarantees (Cohen, 2016).

Leverage score sampling. Let $\mathbf{S}$ be a row sampling sketch of size m, i.e., each row is distributed independently as $\frac{1}{\sqrt{m}}\mathbf{x}^\top$, where $\mathbf{x} = \frac{1}{\sqrt{p_s}}\mathbf{e}_s$ and s is an index drawn from distribution $(p_1, ..., p_n)$. Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d, we call this an approximate leverage score sampling sketch if $p_i \approx_{O(1)} l_i/d$ for all i, where $l_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_i$ is the ith leverage score of $\mathbf{A}$. We will present two lower bound constructions.

1. Approximate sampling and arbitrary $\mathbf{A}$. Now, suppose that n is even and consider the following specific example:

$$p_j = \begin{cases} l_j/2d, & \text{for } j \le n/2, \\ 3l_j/2d, & \text{otherwise.} \end{cases}$$

Further, consider the matrix $\mathbf{B} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{P}$, which is the projection onto the column-span of $\mathbf{A}$, and therefore satisfies the restriction requirement in the Restricted Bai-Silverstein condition. Then, since $\mathrm{tr}(\mathbf{B}^2) = \mathrm{tr}(\mathbf{P}^2) = \mathrm{tr}(\mathbf{P}) = d$, we have:

$$\mathrm{Var}[\mathbf{x}^\top\mathbf{B}\mathbf{x}] = \mathbb{E}\Big[\big(\mathbf{e}_s^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{e}_s/p_s - d\big)^2\Big] = \mathbb{E}\Big[\big(l_s/p_s - d\big)^2\Big] \ge (d/3)^2$$

2. Exact sampling and a specific $\mathbf{A}$. Suppose that $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$, each $\mathbf{a}_i$ is a standard basis vector scaled by $\sqrt{d/n}$ and we are sampling index s according to exact leverage scores, i.e., uniformly at random. Then, letting $\mathbf{x} = \frac{1}{\sqrt{p_s}}\mathbf{e}_s$ and $\mathbf{B} = \mathbf{A}\mathbf{C}\mathbf{A}^\top$, we have:

$$\begin{aligned}\mathrm{Var}[\mathbf{x}^\top\mathbf{B}\mathbf{x}] &= \mathbb{E}\Big[\big(\mathbf{x}^\top\mathbf{B}\mathbf{x} - \mathrm{tr}(\mathbf{B})\big)^2\Big] = \mathbb{E}\Big[\Big(d\cdot\frac{\mathbf{a}_s^\top\mathbf{C}\mathbf{a}_s}{\mathbf{a}_s^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_s} - \mathrm{tr}(\mathbf{C})\Big)^2\Big] \\ &= d^2\cdot\frac{1}{d}\sum_{j=1}^d\Big(C_{jj} - \frac{1}{d}\sum_{i=1}^d C_{ii}\Big)^2 = d^2\cdot\Omega(1), \quad \text{if} \quad C_{ii} = \begin{cases} 1/2, & \text{for even } i, \\ 3/2, & \text{for odd } i. \end{cases}\end{aligned}$$

In both constructions, we have $\mathrm{tr}(\mathbf{B}^2) = \Theta(d)$, so this implies that for leverage score sampling, Condition 2 can only be shown with factor $\alpha = \Omega(d)$, as opposed to $O(1)$ for sub-gaussian sketches and LESS embeddings.

Data-oblivious sparse embeddings. Let $\mathbf{S}$ be a sketch of size m, where each row is distributed independently as $\frac{1}{\sqrt{m}}\mathbf{x}^\top$ and $\mathbf{x} = (\sqrt{\frac{m}{s}}b_1r_1, ..., \sqrt{\frac{m}{s}}b_nr_n)$, with $b_i \sim \mathrm{Bernoulli}(\frac{s}{m})$ and r_i distributed as a uniformly random sign. While this is not the most commonly studied variant of a sparse embedding, it is known to satisfy the subspace embedding property for sketch size $m = O(d\log d)$ with sparsity level $s = O(\log d)$ (Cohen, 2016), which matches the state-of-the-art for sparse embeddings. Other sparse embeddings have non-i.i.d. row distributions (Clarkson and Woodruff, 2017; Nelson and Nguyên, 2013; Meng and Mahoney, 2013), and so they do not fit into the framework laid out by Theorem 11. The key difference between the sparsification of $\mathbf{x}$ relative to our LESS embeddings is that it is data-oblivious. We can exploit that in our lower bound example by choosing an extremely skewed leverage score distribution of matrix $\mathbf{A}$. In particular, suppose that $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ and moreover, $\mathbf{a}_i = \mathbf{e}_i$ for $i = 1, ..., k$ (where $1 \le k \le d$) and for all $i > k$, the first k coordinates of a_i are zero. This construction ensures that the first k leverage scores of $\mathbf{A}$ are equal 1. Once again setting $\mathbf{B} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$, we get:

$$\mathrm{Var}[\mathbf{x}^\top\mathbf{B}\mathbf{x}] \ge \sum_{i=1}^k \mathrm{Var}\Big[\frac{m}{s}b_ir_i\Big] = k\cdot\frac{m}{s}\Big(1-\frac{s}{m}\Big).$$

If we let $k = \Omega(d)$, then we get $\mathrm{Var}[\mathbf{x}^\top\mathbf{B}\mathbf{x}] \ge \Omega(m/s)\cdot\mathrm{tr}(\mathbf{B}^2)$. Thus, unless we zero-out merely a constant fraction of entries of $\mathbf{S}$, the sketching matrix will not satisfy Condition 2 with a constant factor $\alpha = O(1)$. We conjecture that this example can be extended to show a general lower bound on the inversion bias, as we did for approximate leverage score sampling.

Appendix D. Subspace embedding guarantee for LESS embeddings

In this section, we prove Lemma 12 and Theorem 8. In particular, we prove that LESS embeddings achieve the subspace embedding property for a sketch of size $O(d\log d)$ (Lemma 12), thus establishing Condition 1. Then, at the end of the section we briefly discuss how to combine Lemmas 12 and 28, using the structural conditions via Theorem 11, to obtain Theorem 8.

D.1. Proof of Lemma 12

First, note that instead of directly showing the subspace embedding of $\mathbf{SA}$ for the span of $\mathbf{A}$, it suffices to show the guarantee when replacing $\mathbf{A}$ with its orthonormal basis matrix $\mathbf{U} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1/2}$, since $\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A} = (\mathbf{A}^\top\mathbf{A})^{\frac{1}{2}}\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}(\mathbf{A}^\top\mathbf{A})^{\frac{1}{2}}$. Then, a standard technique, e.g., as used for leverage score sampling sketches, relies on the following decomposition of $\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}$ as an average of independent rank-one p.s.d. random matrices:

$$\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U} = \sum_{i=1}^m \mathbf{U}^\top\mathbf{s}_i\mathbf{s}_i^\top\mathbf{U},$$

where $\mathbf{s}_i^\top$ represents the ith row of $\mathbf{S}$. For standard leverage score sampling sketches it suffices to use the matrix Chernoff bound (Tropp, 2012, Theorem 1.1), which uses an almost sure bound on each rank-one matrix to ensure concentration around the mean, $\mathbb{E}[\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}] = \mathbf{I}$. However, in the case of a leverage score sparsified embedding an almost sure bound is not sufficient. Instead, we show that the rank-one matrices $\mathbf{U}^\top \mathbf{s}_i \mathbf{s}_i^\top \mathbf{U}$ exhibit sub-exponential tails on all of their moments, as required by the following variant of the matrix Bernstein bound.

Lemma 29 (Tropp (2012, Theorem 6.2)) *For $i = 1, 2, \ldots$, consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{X}_i] = \mathbf{0}, \quad \mathbb{E}[\mathbf{X}_i^p] \preceq \frac{p!}{2} \cdot R^{p-2} \mathbf{A}_i^2 \quad \textit{for} \quad p = 2, 3, \ldots$$

Then, defining the variance parameter $\sigma^2 = \| \sum_i \mathbf{A}_i^2 \|$, for any $t > 0$ we have:

$$\Pr\Big\{ \lambda_{\max}\Big(\sum\nolimits_i \mathbf{X}_i \Big) \geq t \Big\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt} \right).$$

We apply the above result for $\mathbf{X}_i = \pm(\mathbf{U}^\top \mathbf{s}_i \mathbf{s}_i^\top \mathbf{U} - \frac{1}{m}\mathbf{I})$, where $\mathbf{s}_i = \frac{1}{\sqrt{m}}(\mathbf{x}_i \circ \boldsymbol{\xi})$ is a leverage score sparsified sub-gaussian random vector. We next establish the subexponential moment bound needed for the matrix Bernstein bound.

Lemma 30 *Fix a matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$ such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$. Suppose that $\boldsymbol{\xi}$ is a leverage score sparsifier for $\mathbf{U}$ and $\mathbf{x}$ has i.i.d. $O(1)$-sub-gaussian entries with mean zero and unit variance. Then, there is $C = O(1)$ such that for all $p = 2, 3, \ldots$ we have*

$$\left\| \mathbb{E}\Big[\Big(\mathbf{U}^\top (\mathbf{x} \circ \boldsymbol{\xi})(\mathbf{x} \circ \boldsymbol{\xi})^\top \mathbf{U} - \mathbf{I} \Big)^p \Big] \right\| \leq \frac{p!}{2} \cdot (Cd)^{p-1}.$$

Now, the matrix Bernstein bound (Lemma 29) can be invoked with $\mathbf{A}_i^2 = \frac{Cd}{m^2} \cdot \mathbf{I}$ and $\sigma^2 = R = \frac{Cd}{m}$, obtaining that for $\eta \in (0, 1)$:

$$\Pr\Big\{ \big\| \mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U} - \mathbf{I} \big\| \geq \eta \Big\} \leq 2d \cdot \exp\Big(-\frac{\eta^2 m}{4Cd} \Big) \leq \delta \quad \text{for} \quad m \geq 4Cd \log(2d/\delta)/\eta^2,$$

which completes the proof.

D.2. Proof of Lemma 30

The key part of our proof of Lemma 30 involves establishing the following concentration inequality which can be viewed as a form of the Hanson-Wright inequality (Rudelson and Vershynin, 2013) that takes advantage of the leverage score sparsifier $\boldsymbol{\xi}$, similarly as we did for the Restricted Bai-Silverstein inequality (Lemma 28).

Lemma 31 *Fix a matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$ such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$. Suppose that $\boldsymbol{\xi}$ is a leverage score sparsifier for $\mathbf{U}$ and $\mathbf{x}$ has independent $O(1)$-sub-gaussian entries with mean zero and unit variance. Then, there is $c = \Omega(1)$ and $C = O(1)$ such that for any $t \geq Cd$ we have:*

$$\Pr\Big\{ (\mathbf{x} \circ \boldsymbol{\xi})^\top \mathbf{U} \mathbf{U}^\top (\mathbf{x} \circ \boldsymbol{\xi}) \geq t \Big\} \leq \exp\Big(-c\,\big(\sqrt{t} + t/d\big) \Big).$$

Proof We use the shorthand $\mathbf{U}_{\boldsymbol{\xi}} = \operatorname{diag}(\boldsymbol{\xi})\mathbf{U}$. Similarly as for Lemma 28, our strategy is to show that the sparsification $\mathbf{U}_{\boldsymbol{\xi}}$ preserves enough of the structure of $\mathbf{U}$ so that we can apply the classical Hanson-Wright inequality, which is repeated below, following (Rudelson and Vershynin, 2013),

Lemma 32 (Rudelson and Vershynin (2013, Theorem 1.1)) *Let $\mathbf{x}$ have independent $O(1)$-subgaussian entries with mean zero and unit variance. Then, there is $c = \Omega(1)$ such that for any $n \times n$ matrix $\mathbf{B}$ and $t \geq 0$,*

$$\Pr\Big\{|\mathbf{x}^\top \mathbf{B}\mathbf{x} - \mathrm{tr}(\mathbf{B})| \geq t\Big\} \leq 2\exp\bigg(-c\min\Big\{\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|}\Big\}\bigg).$$

To show that the leverage score sparsification $\mathbf{U}_{\boldsymbol{\xi}}$ is sufficiently accurate, we can rely on the matrix Chernoff bound, repeated below, and the following decomposition:

$$\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}} = \sum_{i=1}^{d} \frac{\mathbf{u}_{s_i}\mathbf{u}_{s_i}^\top}{dp_{s_i}},$$

where s_i are the random indices sampled from the approximate leverage score distribution $(p_1, ..., p_n)$ (see Definition 6). For simplicity, we only repeat the large deviation part of the Chernoff bound, which is the one relevant to our analysis.

Lemma 33 (Tropp (2012, Theorem 1.1 and Remark 5.3)) *For $i = 1, 2, ...$, consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent positive semi-definite random matrices such that $\mathbb{E}\big[\sum_i \mathbf{X}_i\big] = \mathbf{I}$ and $\|\mathbf{X}_i\| \leq R$. Then, for any $t \geq \mathrm{e}$, we have:*

$$\Pr\Big\{\Big\|\sum_i \mathbf{X}_i\Big\| \geq t\Big\} \leq d \cdot \Big(\frac{\mathrm{e}}{t}\Big)^{t/R}.$$

We apply the matrix Chernoff to $\mathbf{X}_i = \frac{1}{dp_{s_i}}\mathbf{u}_{s_i}\mathbf{u}_{s_i}^\top$, noting that since $p_i \geq \|\mathbf{u}_i\|^2/Rd$ for $R = O(1)$, it follows that $\|\mathbf{X}_i\| \leq R$. Moreover, $\mathbb{E}[\sum_{i=1}^d \mathbf{X}_i] = \mathbf{I}$, so for $t \geq O(1) \cdot d$ we have:

$$\Pr\big\{\|\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}\| \geq \sqrt{t}\big\} \leq d\exp\big(-\sqrt{t}\ln(\sqrt{t}/\mathrm{e})/R\big) \leq \exp(-c\sqrt{t}),$$

for some $c = \Omega(1)$. Also, note that $\|\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}\| \leq \mathrm{tr}(\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}) \leq Rd$ almost surely, which implies that event $\mathcal{E} : \big[\|\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}\| \leq \min\{\sqrt{t}, Rd\}\big]$ holds with probability at least $1 - \exp(-c(\sqrt{t} + t/d))$. Conditioned on $\mathcal{E}$, it holds that $\|\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\|_F^2 \leq \|\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}\| \cdot \mathrm{tr}(\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{U}_{\boldsymbol{\xi}}) \leq \min\{\sqrt{t}, Rd\} \cdot Rd$, so applying Lemma 32 for fixed $\boldsymbol{\xi}$ we get:

$$\begin{aligned}
\Pr\big\{\mathbf{x}^\top \mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top \mathbf{x} \geq Rd + t \ \mid\ \boldsymbol{\xi}, \mathcal{E}\big\} &\leq 2\exp\bigg(-c\min\Big\{\frac{t^2}{\|\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\|_F^2}, \frac{t}{\|\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}\|}\Big\}\bigg) \\
&\leq 2\exp\bigg(-c\min\Big\{\frac{t^2}{\min\{\sqrt{t}, Rd\} \cdot Rd}, \frac{t}{\min\{\sqrt{t}, Rd\}}\Big\}\bigg) \\
&\leq 2\exp\big(-c(\sqrt{t} + t/Rd)\big).
\end{aligned}$$

Appropriately rescaling t, we obtain the claim. ■

We are now ready to present the proof of Lemma 30, obtaining subexponential moment bounds for the random matrix $\mathbf{U}^\top(\mathbf{x} \circ \boldsymbol{\xi})(\mathbf{x} \circ \boldsymbol{\xi})^\top \mathbf{U}$, thus completing the proof of the subspace embedding guarantee for leverage score sparsified sketches.

Proof [Proof of Lemma 30] Throughout the proof, we will use the shorthand $\mathbf{U}_{\boldsymbol{\xi}} = \mathrm{diag}(\boldsymbol{\xi})\mathbf{U}$. It is easy to show by induction over p that:

$$\underbrace{\big(\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}-\mathbf{I}\big)^p}_{\mathbf{Z}^p} = \big(\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}-1\big)^{p-1}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}} - \underbrace{\big(\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}-\mathbf{I}\big)^{p-1}}_{\mathbf{Z}^{p-1}}.$$

Thus, it follows that for any $p=2,3,...$ (both even and odd) we have the following upper bound:

$$\big\|\mathbb{E}[\mathbf{Z}^p]\big\| \leq \Big\|\mathbb{E}\big[\underbrace{|\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}-1|^{p-1}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}}_{\mathbf{T}_p}\big]\Big\| + \big\|\mathbb{E}[\mathbf{Z}^{p-1}]\big\|.$$

To bound the quadratic form $\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}$ in the first term, we can use Lemma 31. In particular, the lemma implies that the event $\mathcal{E}\,:\,[\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x} \leq Cpd]$ fails with probability at most $\mathrm{e}^{-\sqrt{pd}}$ for a sufficiently large $C=O(1)$, so we have:

$$\begin{aligned}
\big\|\mathbb{E}[\mathbf{T}_p]\big\| &\leq \big\|\mathbb{E}[\mathbf{T}_p\cdot\mathbf{1}_{\mathcal{E}}]\big\| + \big\|\mathbb{E}[\mathbf{T}_p\cdot\mathbf{1}_{\neg\mathcal{E}}]\big\| \\
&\leq (pd)^{p-1}\big\|\mathbb{E}[\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}]\big\| + \mathbb{E}\big[(\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\cdot\mathbf{1}_{\neg\mathcal{E}})^p\big] \\
&= (Cpd)^{p-1} + \int_0^\infty pt^{p-1}\mathrm{Pr}\big\{\mathbf{x}^\top\mathbf{U}_{\boldsymbol{\xi}}\mathbf{U}_{\boldsymbol{\xi}}^\top\mathbf{x}\cdot\mathbf{1}_{\neg\mathcal{E}} > t\big\}dt \\
&\leq (Cpd)^{p-1} + p(Cpd)^p\mathrm{e}^{-\sqrt{pd}} + \int_{Cpd}^\infty pt^{p-1}\mathrm{e}^{-c(\sqrt{t}+t/d)}dt.
\end{aligned}$$

Note that $(O(1)\,p)^{p+O(1)}d^{p-1} \leq p^p(O(1)\,d)^{p-1} \leq (p!/2)(O(1)\,d)^{p-1}$, and also $\mathrm{e}^{-\sqrt{pd}} \leq O(1/d)$, so the first two terms can be easily bounded as desired. To bound the last term, we use the following integral formula:

$$\int t^{p-1}\mathrm{e}^{-\alpha t^\theta}dt = -\frac{\Gamma(p/\theta,\alpha t^\theta)}{\theta\alpha^{p/\theta}} + \mathrm{const},$$

which follows from the definition of the upper incomplete Gamma function Γ. Note that for $p=2,3,...$ this function also satisfies:

$$\begin{aligned}
\Gamma(p,\lambda) &= (p-1)!\cdot\mathrm{Pr}\{x<p\} \qquad \text{for} \quad x\sim\mathrm{Poisson}(\lambda), \\
&\leq (p-1)!\cdot\mathrm{e}^{-c\lambda} \qquad \text{for} \quad \lambda\geq 2p,\ c=\Omega(1),
\end{aligned}$$

where the last inequality is a standard tail bound for a Poisson random variable. With a slight abuse of notation, we let c denote the minimum of the above constant c and the constant c from Lemma 31. We apply the integral formula in two different ways, depending on p. First, if $p<d$ then we have:

$$\int_{Cpd}^\infty pt^{p-1}\mathrm{e}^{-c(\sqrt{t}+t/d)}dt \leq \int_{Cpd}^\infty pt^{p-1}\mathrm{e}^{-c\sqrt{t}}dt = 2pc^{-2p}\Gamma(2p,c\sqrt{Cpd}) \leq 2c^{-2p}(2p)!\mathrm{e}^{-c^2\sqrt{Cpd}}.$$

By using the fact that $\exp(-c^2\sqrt{Cpd}) \leq \exp(-c^2p) = O(1/p)$, this expression can be bounded by $(p!/2)(O(1)\,p)^{p-1} \leq (p!/2)(O(1)\,d)^{p-1}$. Next, we consider the case when $p\geq d$. We have:

$$\int_{Cpd}^\infty pt^{p-1}\mathrm{e}^{-c(\sqrt{t}+t/d)}dt \leq \int_{Cpd}^\infty pt^{p-1}\mathrm{e}^{-ct/d}dt = p(d/c)^p\Gamma(p,cCp) \leq p!d^p\mathrm{e}^{-c^2Cp},$$

where the last inequality holds as long as $C \geq 2/c$. Here, we note that $\mathrm{e}^{-c^2 Cp} \leq O(1/d)$ since $p \geq d$, thus again obtaining a bound of the form $(p!/2)(O(1)\,d)^{p-1}$. Putting everything together, we conclude that:

$$\|\mathbb{E}[\mathbf{Z}^p]\| \leq \frac{p!}{2}\,(O(1)\,d)^{p-1} + \|\mathbb{E}[\mathbf{Z}^{p-1}]\|.$$

Recursively summing up this bound concludes the proof. ■

D.3. Proof of Theorem 8

In Lemma 12, we showed that a LESS embedding of size $m \geq 4Cd\log(3d/\delta)$ satisfies Condition 1 (subspace embedding) for $\eta = 1/2$ with probability $1-\delta/3$, as required by Theorem 11. Also, in Lemma 28 we showed Condition 2 (Restricted Bai-Silverstein) with $\alpha = O(1)$ for a leverage score sparsified sub-gaussian vector. Thus, as long as $\delta \leq 1/m^3$ and $m/3 \geq 4Cd\log(3d/\delta)$, it follows that $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an (ϵ,δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ for $\epsilon = O(\sqrt{d}/m)$, and we obtain the desired guarantee. Note that the condition for invoking Theorem 11 can be written as $m \geq C' d\log(m)$. This is satisfied for $m = C' d\log(C'^2 d^2)$, and since m grows faster than $\log(m)$, it will also be satisfied for all $m \geq C' d\log(C'^2 d^2) = O(d\log(d))$. This completes the proof of Theorem 8.

Appendix E. Averaging nearly-unbiased estimators

In this section, we show that averaging improves spectral approximation for matrix estimators with small inversion bias, and as a consequence we prove Corollaries 5 and 9 for averaging sketched inverse covariance matrix estimators based on sub-gaussian sketches and LESS embeddings respectively.

E.1. Conditions for effective averaging of random matrices

We start with a more general result, which should be of interest to averaging nearly-unbiased matrix estimators in settings other than inverse covariance matrix estimation.

Lemma 34 (Conditions for effective averaging) *Suppose that $\delta \leq \epsilon \leq \eta \leq 1$ and $\tilde{\mathbf{C}}_1, ..., \tilde{\mathbf{C}}_q$ are i.i.d. positive semi-definite d-dimensional random matrices such that:*

1. *$\tilde{\mathbf{C}}_i$ is an $(\epsilon, \delta/2q)$-unbiased estimator of $\mathbf{C}$;*
2. *$\tilde{\mathbf{C}}_i$ is an $(\eta, \delta/2q)$-approximation of $\mathbf{C}$.*

Then, $\frac{1}{q}\sum_{i=1}^q \tilde{\mathbf{C}}_i$ is an $(\epsilon', 2\delta)$-approximation of $\mathbf{C}$ for $\epsilon' = \epsilon + \eta \cdot O\big(\sqrt{\frac{\ln(d/\delta)}{q}}\big)$.

Proof For this, we use a variant of the matrix Bernstein inequality given below.

Lemma 35 (Tropp (2012, Theorem 1.4)) *For $i = 1, 2, ...$, consider a finite sequence $\mathbf{X}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{X}_i] = \mathbf{0}, \quad \lambda_{\max}(\mathbf{X}_i) \leq R \quad \textit{almost surely.}$$

Then, defining the variance parameter $\sigma^2 = \|\sum_i \mathbb{E}[\mathbf{X}_i^2]\|$, for any $t > 0$ we have:

$$\Pr\Big\{\lambda_{\max}\Big(\sum\nolimits_i \mathbf{X}_i\Big) \geq t\Big\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Suppose that $\tilde{\mathbf{C}}$ is an $(\epsilon, \delta/2q)$-unbiased estimator and an $(\eta, \delta/2q)$-spectral approximation for $\mathbf{C}$, with $\mathcal{E}_{inv}$ and $\mathcal{E}_{sub}$ the associated high probability events. For concreteness, let the $O(1)$ constant factor in Definition 3 be denoted as M. Further, let $\tilde{\mathbf{C}}_i, \mathcal{E}_{inv}^i, \mathcal{E}_{sub}^i$ be the i.i.d. copies of $\tilde{\mathbf{C}}$ with their associated events. Finally, let $\tilde{\mathbf{C}}_i'$ be a random matrix obtained from conditioning $\tilde{\mathbf{C}}_i$ on $\mathcal{E}_{inv}^i \wedge \mathcal{E}_{sub}^i$, and coupled with $\tilde{\mathbf{C}}_i$ so that $\Pr(\tilde{\mathbf{C}}_i' = \tilde{\mathbf{C}}_i) \geq \Pr(\mathcal{E}_{inv}^i \wedge \mathcal{E}_{sub}^i) \geq 1 - \delta/q$ (this coupling can be obtained by considering a construction of $\tilde{\mathbf{C}}_i'$ via rejection sampling from $\tilde{\mathbf{C}}_i$). We can bound the bias of $\tilde{\mathbf{C}}_i'$ (for any i) by observing that:

$$-\delta/q \cdot \mathbb{E}[\tilde{\mathbf{C}}_i \mid \mathcal{E}_{inv}^i, \neg\mathcal{E}_{sub}^i] \preceq \mathbb{E}[\tilde{\mathbf{C}}_i'] - \mathbb{E}[\tilde{\mathbf{C}}_i \mid \mathcal{E}_{inv}^i] \preceq \frac{\delta/q}{1-\delta/q} \cdot \mathbb{E}[\tilde{\mathbf{C}}_i \mid \mathcal{E}_{inv}^i].$$

Since we have $\mathbb{E}[\tilde{\mathbf{C}}_i \mid \mathcal{E}_{inv}^i] \approx_\epsilon \mathbf{C}$ and $\mathbb{E}[\tilde{\mathbf{C}}_i \mid \mathcal{E}_{inv}^i, \neg\mathcal{E}_{sub}^i] \preceq M \cdot \mathbf{C}$, it follows that $\mathbb{E}[\tilde{\mathbf{C}}_i']$ is an ϵ'-spectral approximation of $\mathbf{C}$ for $\epsilon' = \epsilon + \frac{2\delta}{q}(1 + \epsilon + M)$.

We will now apply the matrix Bernstein inequality (Lemma 35) to the sequence of matrices:

$$\mathbf{X}_i = \frac{1}{q}\Big(\mathbf{C}^{-\frac{1}{2}}\tilde{\mathbf{C}}_i'\mathbf{C}^{-\frac{1}{2}} - \mathbb{E}\big[\mathbf{C}^{-\frac{1}{2}}\tilde{\mathbf{C}}_i'\mathbf{C}^{-\frac{1}{2}}\big]\Big), \quad i = 1, ..., q.$$

Note that we have $\tilde{\mathbf{C}}_i' \approx_\eta \frac{1}{q}\mathbf{C}$, so it follows that $\|\mathbf{X}_i\| \leq (\eta + \epsilon')/q$ and $\sum_i \|\mathbf{X}_i^2\| \leq (\eta + \epsilon')^2/q$. Thus, we conclude that for $t \in (0, 1)$:

$$\Pr\Big\{\Big\|\sum_{i=1}^q \mathbf{X}_i\Big\| \geq t\,(\eta + \epsilon')\Big\} \leq 2d\exp\big(-t^2 q/4\big).$$

Setting $t = \sqrt{4\ln(2d/\delta)/q}$ (without loss of generality, assume that $t \leq 1$), we obtain that with probability $1 - \delta$,

$$\begin{aligned}\Big\|\frac{1}{q}\sum_{i=1}^q \mathbf{C}^{-\frac{1}{2}}\tilde{\mathbf{C}}_i'\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\Big\| &\leq \Big\|\sum_{i=1}^q \mathbf{X}_i\Big\| + \Big\|\frac{1}{q}\sum_{i=1}^q \mathbf{C}^{-\frac{1}{2}}\mathbb{E}[\tilde{\mathbf{C}}_i']\mathbf{C}^{-\frac{1}{2}} - \mathbf{I}\Big\| \\ &\leq t \cdot (\eta + \epsilon') + \epsilon' \leq \epsilon + \eta \cdot O\Big(\sqrt{\tfrac{\log(d/\delta)}{q}}\Big) + O\Big(\frac{\delta M}{q}\Big).\end{aligned}$$

Note that under the assumptions that $M = O(1)$ and $\delta \leq \eta$, we can absorb the last term into the middle term. Finally, observe that thanks to the coupling and a union bound, the above bound holds with probability $1 - 2\delta$ after we replace $\tilde{\mathbf{C}}_i'$ with $\tilde{\mathbf{C}}_i$, completing the proof of Lemma 34. ■

E.2. Proof of Corollary 5

Consider a sub-gaussian sketching matrix $\mathbf{S}$ of size $m \geq C(d + \sqrt{d}/\epsilon + \log(2q/\delta))$. From Proposition 4, it follows that $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an $(\epsilon, \delta/2q)$-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$. Further, it is an $(\eta, \delta/2q)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$, where $\eta = O(\sqrt{d/m}) = O(\epsilon \cdot \sqrt{m})$. Thus, using

Lemma 34, it follows that for q i.i.d. copies $\mathbf{S}_1, ..., \mathbf{S}_q$, the averaged estimator $\frac{1}{q}\sum_{i=1}^q(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_i^\top\mathbf{S}_i\mathbf{A})^{-1}$ is an $(\epsilon'', 2\delta)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$ for

$$\epsilon'' = \epsilon + O\Big(\epsilon \cdot \sqrt{m\log(d/\delta)/q}\Big).$$

Setting $q = O(m\log(d/\delta))$ and adjusting the constants appropriately, we obtain the claim.

E.3. Proof of Corollary 9

Consider a LESS embedding matrix $\mathbf{S}$ of size $m \geq C(d\log(2dq/\delta) + \sqrt{d}/\epsilon)$. From Theorem 8, it follows that $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an $(\epsilon, \delta/2q)$-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$. Furthermore, the theorem also implies that this matrix is an $(\eta, \delta/2q)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$ for $\eta = O(\sqrt{d\log(2dq/\delta)/m}) = O(\epsilon \cdot \sqrt{m\log(d/\delta)})$. Using Lemma 34, it follows that for q i.i.d. copies $\mathbf{S}_1, ..., \mathbf{S}_q$, the averaged estimator $\frac{1}{q}\sum_{i=1}^q(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}_i^\top\mathbf{S}_i\mathbf{A})^{-1}$ is an $(\epsilon'', 2\delta)$-approximation of $(\mathbf{A}^\top\mathbf{A})^{-1}$ for

$$\epsilon'' = \epsilon + O\Big(\epsilon \cdot \sqrt{m\log^2(2dq/\delta)/q}\Big).$$

Setting $q = O(m\log^2(d/\delta))$ and adjusting the constants appropriately, we obtain the claim.

Note that in both Corollaries there is a slight interdependence in the conditions for m and q. This is in general unavoidable, since as q grows large with fixed m, the average has to eventually converge to the true expectation of $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$, which may be unbounded.

Appendix F. Inversion bias lower bound for leverage score sampling

In this section, we show a lower bound on the inversion bias of approximate leverage score sampling, proving Theorem 10. In the proof, we show a lower bound for the inverse moment of a shifted Binomial random variable (Lemma 36), which should be of independent interest.

F.1. Proof of Theorem 10

Without loss of generality, suppose that $n = 2d$ (otherwise the matrix $\mathbf{A}$ can be padded by zeros). We can also assume that $m \geq d$, since the other cases follow easily. Our construction is designed so that uniform row sampling is a 1/2-approximation of leverage score sampling. Let $\mathbf{S}$ be a uniform row sampling sketch of size m, i.e., its ith row is $\sqrt{\frac{n}{m}}\,\mathbf{e}_{s_i}^\top$, where $s_1, ..., s_m$ are independent uniformly random indices from $1, ..., n$. Our matrix $\mathbf{A}$ consists of $n = 2d$ scaled standard basis vectors such that pairs of consecutive rows are given by $\mathbf{a}_{2(i-1)+1}^\top = \mathbf{a}_{2(i-1)+2}^\top = \frac{1}{\sqrt{2}}\,\mathbf{e}_i^\top$ for $i \geq 2$, whereas the first two rows are $\mathbf{a}_1^\top = \frac{1}{\sqrt{4}}\mathbf{e}_1^\top$ and $\mathbf{a}_2^\top = \sqrt{\frac{3}{4}}\mathbf{e}_1^\top$:

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{4}} & & & \\ \sqrt{\frac{3}{4}} & & 0 & \\ & \frac{1}{\sqrt{2}} & & \\ & \frac{1}{\sqrt{2}} & & \\ & & \ddots & \\ & 0 & & \frac{1}{\sqrt{2}} \\ & & & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

First, note that $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$, and all of the squared row norms are within $[\frac{1}{2}\frac{d}{n}, \frac{3}{2}\frac{d}{n}]$, so uniform sampling is indeed a $1/2$-approximate leverage score sampling scheme. Further, for any $\gamma > 0$, the matrix $\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}$ is diagonal, and its diagonal entries are given by:

$$\big[\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}\big]_{ii} = \begin{cases} \frac{\gamma n}{m}\sum_{j=1}^m \left(\frac{1}{4}\mathbf{1}_{[s_j=1]} + \frac{3}{4}\mathbf{1}_{[s_j=2]}\right) = \frac{\gamma n}{m}\cdot\frac{x+b_1/2}{2} & \text{for } i=1, \\ \frac{\gamma n}{m}\sum_{j=1}^m \left(\frac{1}{2}\mathbf{1}_{[s_j=2(i-1)+1]} + \frac{1}{2}\mathbf{1}_{[s_j=2(i-1)+2]}\right) = \frac{\gamma n}{m}\cdot\frac{b_i}{2} & \text{otherwise,} \end{cases}$$

where b_i's are all identically (but not independently) distributed as $\mathrm{Binomial}(m, 1/d)$ and x is distributed, conditionally on b_1, as $\mathrm{Binomial}(b_1, 1/2)$. Here b_i denote the number of times $s_j \in \{2i-1, 2i\}$, while x denotes the number of times $s_j = 2$. Due to the symmetry of the problem, conditionally on a given value of b_1 (i.e., a given value of counts s_j that are equal to either unity or two), each $s_j \in \{1,2\}$ is distributed uniformly over $\{1,2\}$, hence the value x of counts s_j that are equal to two is distributed as $\mathrm{Binomial}(b_1, 1/2)$. This leads to the claimed distributional representation.

The key idea in the construction is that the first diagonal entry of the sketch has more variance than the others, and thus it will also have more inversion bias. As a result, there is no scaling γ that will simultaneously correct the inversion bias of the first entry and of all the other entries. To that end, we lower bound a shifted inverse moment of the Binomial distribution in the following lemma, potentially of independent interest, proven at the end of this section.

Lemma 36 *There is a universal constant $C > 0$ such that for any positive integer b, if $x \sim \mathrm{Binomial}(b, 1/2)$ then:*

$$\mathbb{E}\bigg[\frac{1}{x + b/2}\bigg] \geq \Big(1 + \frac{1}{Cb}\Big)\cdot\frac{1}{b}.$$

Note that the expected inverse of $\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A}$ is undefined since the matrix may not be invertible. Thus, as in the definition of an (ϵ,δ)-unbiased estimator, we must condition on a high probability event which ensures invertibility. We start by considering the largest such event, $\mathcal{E}^* : [\forall_i b_i > 0]$.

Using the fact that, conditioned on b_1, the variable x is independent of $\mathcal{E}^*$, we have:

$$\begin{aligned}
\mathbb{E}\Big[\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{11} \mid \mathcal{E}^*\Big] &= \Big(\frac{\gamma n}{m}\Big)^{-1}\sum_{b>0}\mathbb{E}\Big[\frac{2}{x+b_1/2} \mid b_1=b\Big]\Pr(b_1=b\mid\mathcal{E}^*)\\
&\overset{(a)}{\geq} \Big(\frac{\gamma n}{m}\Big)^{-1}\sum_{b>0}\Big(1+\frac{1}{Cb}\Big)\frac{2}{b}\Pr(b_1=b\mid\mathcal{E}^*)\\
&= \sum_{b>0}\Big(1+\frac{1}{Cb}\Big)\mathbb{E}\Big[\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{22}\mid b_2=b\Big]\Pr(b_2=b\mid\mathcal{E}^*)\\
&\geq \mathbb{E}\Big[\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{22}\mid\mathcal{E}^*\Big]\\
&\quad+\frac{1}{2C}\frac{d}{m}\sum_{b=1}^{2m/d}\mathbb{E}\Big[\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{22}\mid b_2=b\Big]\Pr(b_2=b\mid\mathcal{E}^*)\\
&\overset{(b)}{\geq}\Big(1+\frac{d}{4Cm}\Big)\cdot\mathbb{E}\Big[\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{22}\mid\mathcal{E}^*\Big],
\end{aligned}$$

where in (a) we used Lemma 36 and in (b) we observed that $\big[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\big]_{22}$ decreases with b_2 and moreover, since $\mathbb{E}[b_2]=m/d\geq 1$, it is easy to verify that the range $[1,2m/d]$ contains more than half of the probability mass of $\mathrm{Binomial}(m,1/d)$.

The above derivation shows that when conditioned on $\mathcal{E}^*$, for any scaling $\gamma>0$ the inversion bias will be at least $\Omega(d/m)$, since the estimated matrix $(\mathbf{A}^\top\mathbf{A})^{-1}=\mathbf{I}$ has the same entries on the diagonal, whereas the expectation of the first two diagonal entries of the estimator $(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ differs by a factor of $1+\Omega(d/m)$. To complete the proof of Theorem 10, it remains to show that the same is true not just for $\mathcal{E}^*$, but for any event $\mathcal{E}\subseteq\mathcal{E}^*$ with sufficiently high probability. Suppose that $\mathcal{E}$ is such an event, with $\delta=\Pr(\mathcal{E}\mid\mathcal{E}^*)\leq\Pr(\neg\mathcal{E})\leq\frac{1}{4C\cdot 16}(\frac{d}{m})^2$. Then, using $\tau_i=\frac{m}{\gamma n}[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}]_{ii}$ as a shorthand, we have:

$$\mathbb{E}[\tau_1\mid\mathcal{E}]=\mathbb{E}[\tau_1\mid\mathcal{E}^*]+\frac{\delta}{1-\delta}\Big(\mathbb{E}[\tau_1\mid\mathcal{E}^*]-\mathbb{E}[\tau_1\mid\mathcal{E}^*,\neg\mathcal{E}]\Big)\geq\mathbb{E}[\tau_1\mid\mathcal{E}^*]-8\delta,$$

where we used that $\delta\leq 1/2$ and, conditioned on $\mathcal{E}^*$, we have $\tau_1\leq 4$. On the other hand,

$$\mathbb{E}[\tau_2\mid\mathcal{E}]=\mathbb{E}[\tau_2\mid\mathcal{E}^*]+\frac{\delta}{1-\delta}\Big(\mathbb{E}[\tau_2\mid\mathcal{E}^*]-\mathbb{E}[\tau_2\mid\mathcal{E}^*,\neg\mathcal{E}]\Big)\leq(1+2\delta)\,\mathbb{E}[\tau_2\mid\mathcal{E}^*].$$

Combining the two inequalities and using that $\mathbb{E}[\tau_2\mid\mathcal{E}^*]\geq d/m$ and $\delta\leq\frac{1}{4C\cdot 16}(\frac{d}{m})^2$, we get:

$$\begin{aligned}
\frac{\mathbb{E}\big[[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}]_{11}\mid\mathcal{E}\big]}{\mathbb{E}\big[[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}]_{22}\mid\mathcal{E}\big]}&=\frac{\mathbb{E}[\tau_1\mid\mathcal{E}]}{\mathbb{E}[\tau_2\mid\mathcal{E}]}\geq\frac{(1+\frac{d}{4Cm})\mathbb{E}[\tau_2\mid\mathcal{E}^*]-8\delta}{(1+2\delta)\mathbb{E}[\tau_2\mid\mathcal{E}^*]}\\
&\geq\frac{1+\frac{d}{4Cm}-8\delta\frac{m}{d}}{1+2\delta}\geq\frac{1+\frac{d}{8Cm}}{1+\frac{d}{32Cm}}\geq 1+\frac{d}{64Cm}.
\end{aligned}$$

Thus, as discussed above, we conclude that for any scaling $\gamma>0$ and any event $\mathcal{E}$ with probability $\Pr(\mathcal{E})\geq 1-\frac{1}{4C\cdot 16}(\frac{d}{m})^2$, we have $\|\mathbb{E}[(\gamma\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\mid\mathcal{E}]-\mathbf{I}\|=\Omega(\frac{d}{m})$, which concludes the proof.

F.2. Proof of Lemma 36

We conclude this section with a proof of the Binomial inverse moment bound from Lemma 36. While existing work has focused on asymptotic expansions of inverse moments of the Binomial (Znidaric, 2009), those precise characterizations either break down or appear to be impractical to work with when the variable is significantly shifted, as in our case. Thus, we use a different strategy: reducing the inverse moment bound to showing an anti-concentration inequality for the Binomial distribution. For this, we use the classical Paley-Zygmund inequality, stated below.

Lemma 37 *For any non-negative variable* $\mathbf{Z}$ *with finite variance and* $\theta \in (0,1)$*, we have:*

$$\Pr\big(Z \geq \theta\,\mathbb{E}[Z]\big) \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

Let $x \sim \mathrm{Binomial}(b, 1/2)$ for a positive integer b. It follows that:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{x+b/2} - \frac{1}{b}\Big] &= \sum_{i=0}^{b} \Pr(x=i)\Big(\frac{1}{i+b/2} - \frac{1}{b}\Big) = \frac{1}{b}\sum_{i=0}^{b}\Pr(x=i)\frac{b/2-i}{b/2+i} \\ &= \frac{1}{b}\sum_{i=0}^{\lfloor b/2 \rfloor} \Pr(x=i)(b/2-i)\Big(\frac{1}{b/2+i} - \frac{1}{3b/2-i}\Big),\end{aligned}$$

where the last equality is obtained by symmetrically pairing up the terms i and $b-i$ in the first sum. Next, observe that for $0 \leq i \leq b/2 - \sqrt{b}/4$, we have:

$$(b/2-i)\Big(\frac{1}{b/2+i} - \frac{1}{3b/2-i}\Big) \geq \frac{\sqrt{b}}{4}\Big(\frac{1}{b-\sqrt{b}/4} - \frac{1}{b+\sqrt{b}/4}\Big) = \frac{\sqrt{b}}{4}\cdot\frac{\sqrt{b}/2}{b^2-b/16} \geq \frac{1}{8b}.$$

Putting this together, we conclude that:

$$\mathbb{E}\Big[\frac{1}{x+b/2}\Big] \geq \Big(1 + \frac{1}{8b}\Pr\{x - b/2 \leq -\sqrt{b}/4\}\Big)\cdot\frac{1}{b}. \tag{8}$$

Thus, it suffices to show that, with constant probability, x is smaller than its mean, $b/2$, by at least $\sqrt{b}/4$. This follows from the Paley-Zygmund inequality (Lemma 37) by setting $\mathbf{Z} = (x-b/2)^2$. Using standard formulas for the second and fourth centered moment of the Binomial distribution, we have $\mathbb{E}[Z] = b/4$ and $\mathbb{E}[Z^2] = \frac{b}{4}(1 + \frac{3b-6}{4}) \leq 3b^2/16$. Therefore, setting $\theta = 1/4$ in Lemma 37, we obtain:

$$\begin{aligned}\Pr\big(x - b/2 \,\leq\, -\sqrt{b}/4\big) &= \frac{1}{2}\Pr\big(|x-b/2| \,\geq\, \sqrt{b}/4\big) = \frac{1}{2}\Pr\big(Z \geq \theta\,\mathbb{E}[Z]\big) \\ &\geq \frac{1}{2}\Big(1 - \frac{1}{4}\Big)^2 \frac{b^2/16}{3b^2/16} = \frac{3}{32}.\end{aligned}$$

Combining this with (8), we obtain the desired claim for $C = 8\cdot 32/3$.

Appendix G. Exact bias-correction for orthogonally invariant embeddings

In this section we prove that orthogonal invariance implies no inversion bias. This claim has been mentioned in the main text, in Section 2. Here we give a formal statement.

Proposition 38 (Orthogonal invariance implies no inversion bias) *Let $\mathbf{S}$ be a random and right-orthogonally invariant matrix; specifically an $m \times n$ matrix (with $m \leq n$) such that for any orthogonal $n \times n$ matrix $\mathbf{O}$, we have $\mathbf{S} \overset{d}{=} \mathbf{SO}$. Assume that $(\mathbf{A}^\top \mathbf{S}^\top \mathbf{SA})^{-1}$ exists with probability one. Then the inversion bias is exactly correctable, i.e., there exists a constant $c = c_{m,n,d}$ such that $\mathbb{E}\hat{\mathbf{\Sigma}}^{-1} = c \cdot \mathbf{\Sigma}^{-1}$; where $\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{A}$ and $\hat{\mathbf{\Sigma}} = \mathbf{A}^\top \mathbf{S}^\top \mathbf{SA}$.*

Examples of orthogonal ensembles can be constructed in the following way:

1. Let $\mathbf{S}$ have i.i.d. normal entries with variance m^{-1}. Due to the properties of the Wishart ensemble, the constant $c_{m,n,d}$ is $c_{m,n,d} = m/(m-d-1)$.

2. Let $\mathbf{S}_u$ be a uniformly random $m \times n$ partial orthogonal matrix (with $m \leq n$) such that $\mathbf{S}_u \mathbf{S}_u^\top = \mathbf{I}_m$. Equivalently, these are the first few rows of a Haar matrix. Then define $\mathbf{S} = \sqrt{n/m} \cdot \mathbf{S}_u$, scaled such that $\mathbb{E}\mathbf{S}^\top \mathbf{S} = \mathbf{I}_m$. We will call this the Haar sketch.

3. The class of orthogonally invariant matrices has several closure properties. Specifically, it is closed with respect to left-multiplication by any matrices, right-multiplication by orthogonal matrices, and with respect to vector space operations (addition and multiplication by scalars). Several examples can be obtained this way. For instance, matrices $\mathbf{S}$ of the form $\mathbf{S} = \mathbf{MZ}$, where $\mathbf{Z}$ has i.i.d. normal entries with variance m^{-1}, and $\mathbf{M}$ is an arbitrary matrix fixed or random and independent of $\mathbf{Z}$ are orthogonally invariant.

Proof We start with a reduction to orthogonal matrices: Let $\mathbf{A} = \mathbf{U\Lambda V}^\top$ be the SVD of $\mathbf{A}$. Here recall that $\mathbf{A}$ is an $n \times d$ matrix, with $n \geq d$ and with full column rank, and thus $\mathbf{U}$ is an $n \times d$ partial orthogonal matrix with $n \geq d$, $\mathbf{\Lambda}$ is $d \times d$ diagonal, and $\mathbf{V}$ is $d \times d$ orthogonal. Our goal is to show that $\mathbb{E}\hat{\mathbf{\Sigma}}^{-1} = c \cdot \mathbf{\Sigma}^{-1}$, or equivalently that

$$\mathbb{E}(\mathbf{V\Lambda U}^\top \mathbf{S}^\top \mathbf{SU\Lambda V}^\top)^{-1} = c \cdot (\mathbf{V\Lambda U}^\top \mathbf{U\Lambda V}^\top)^{-1}.$$

Then, by cancelling $\mathbf{\Lambda}$ and $\mathbf{V}$ above (using that they are deterministic square invertible matrices), and using that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, we see that the above inequality is equivalent to

$$\mathbb{E}(\mathbf{U}^\top \mathbf{S}^\top \mathbf{SU})^{-1} = c \cdot \mathbf{I}.$$

Thus, the problem is reduced to studying orthogonal matrices $\mathbf{A} = \mathbf{U}$, such that $\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{A} = \mathbf{U}^\top \mathbf{U} = \mathbf{I}$.

We claim that the right-orthogonal invariance implies that $\mathbf{SU} \overset{d}{=} \mathbf{SUO}$ for any $d \times d$ orthogonal matrix $\mathbf{O}$. Here is a geometric argument. We have that $\mathbf{SU}$ are the angles that the random orthogonal rows of $\mathbf{S}$ form with the fixed set of basis vectors formed by the columns of $\mathbf{U}$. Also, $\mathbf{SUO}$ corresponds to the same quantity, but with respect to the basis formed by $\mathbf{UO}$. Since $\mathbf{S}$ is right-rotationally invariant, these angles have the same distribution.

Another, more algebraic proof is as follows. Since $\mathbf{S}$ is right-rotationally invariant, for any orthogonal $n \times n$ matrix $\mathbf{R}$, we have $\mathbf{S} \overset{d}{=} \mathbf{SR}$. Thus, for any fixed matrix $\mathbf{U}$, we have $\mathbf{SU} \overset{d}{=} \mathbf{SRU}$.

Choose a rotation matrix $\mathbf{R}$ such that $\mathbf{R}\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{O}\mathbf{U}^\top$, while $\mathbf{R}\mathbf{U}^\perp$ is arbitrary, where $\mathbf{U}^\perp$ is an orthogonal complement of $\mathbf{U}$. Then, multiplying the above with $\mathbf{U}^\top\mathbf{U}$ we have

$$\mathbf{S}\mathbf{U}\mathbf{U}^\top\mathbf{U} \overset{d}{=} \mathbf{S}\mathbf{R}\mathbf{U}\mathbf{U}^\top\mathbf{U} = \mathbf{S}\mathbf{U}\mathbf{O}\mathbf{U}^\top\mathbf{U} = \mathbf{S}\mathbf{U}\mathbf{O}.$$

We get that $\mathbf{S}\mathbf{U} \overset{d}{=} \mathbf{S}\mathbf{U}\mathbf{O}$. Next, $\mathbf{S}\mathbf{U} \overset{d}{=} \mathbf{S}\mathbf{U}\mathbf{O}$ implies that, with $\mathbf{J} := \mathbb{E}(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}$,

$$\begin{aligned}
\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U} &\overset{d}{=} \mathbf{O}^\top\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}\mathbf{O}\\
(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1} &\overset{d}{=} \mathbf{O}^\top(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}\mathbf{O}\\
\mathbb{E}(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1} &= \mathbf{O}^\top\mathbb{E}(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}\mathbf{O}\\
\mathbf{J} &= \mathbf{O}^\top\mathbf{J}\mathbf{O}.
\end{aligned}$$

Since $\mathbf{J}$ is preserved under conjugation by any orthogonal matrix, $\mathbf{J}$ must be a multiple of the identity matrix, so $\mathbf{J} = c\mathbf{I}_d$, for some $c = c_{m,n,d}$. This finishes the proof. ■